Numerical Mathematics (Numerics of ODEs)

Prof. Ralf Hiptmair, Dr. Vasile Gradinaru, Prof. Philipp Grohs

Seminar for Applied Mathematics, ETH Zürich

P. Grohs

rev 63606, February 17, 2014

Subversion rev 63606

http://www.math.ethz.ch/~pgrohs/tmp/NUMODE16.pdf

Contents

General Informationen

e Lecturer:	Prof. Philipp Grohs, SAM, D-MATH,	Office: HG G 59.2	Tel.: 044 632 32 00, pgrohs@sam.math.ethz.ch
Assistant:	Wei Wu, SAM, D-MATH	Office: HG J 47,	Tel.: 044 632 4320, wei.wu@sam.math.ethz.ch
	Zeljko Kereta, SAM, D-MATH	Office: HG G 56.2,	Tel.: 044 632 2357, zeljko.kereta@sam.math.ethz.c
	Markus Sprecher, SAM, D-MATH	Office: HG J 48,	Tel.: 044 632 6965, markus.sprecher@sam.math.e
TAs:	François Chalus		chalusf@student.ethz.ch
	Mariia Dobrynina		mariiad@student.ethz.ch
	Michael Näf		naefmi@student.ethz.ch

Website:

`http://www.math.ethz.ch/education/bachelor/lectures/fs2016/math/nm2`

Exam: There will be two short exams during semester called "Mid-term exam" and "End-term exam". Both will be worth 12 marks and consist of only routine problems. A student doesn't have to take part in these exams, but those who take part in and do well in these two exams can get bonus in final grade. Suppose Student Z gets x points in Mid-term exam and y points in End-term exam, then student Z will get Max$\{(x+y-12)/2, 0\}$ points of bonus in the final grade.

The final exam will be a written exam *with computer* (partly MATLAB-based programming problems), *no* (brought-along) auxiliary tools

Course notes will be provided as pdf-file

Homework:

- Weekly problem sheet available for download (to be handed in within 1 week)
- MATLAB-based programming problems
- Deadline: Tuesdays until 17:00 am in the boxes in the anteroom of HG G 53
- Codes should be handed in via web-upload: `http://www.math.ethz.ch/~grsam/submit/`
- As testate conditions are not in place anymore, it is not compulsory, yet it is recommended, to hand in the exercises for correction.

Office hours of assistants: will be announced soon on the website of the course.

☞ Literature:

P. DEUFLHARD AND F. BORNEMANN, *Numerische Mathematik II*, DeGruyter, Berlin, 2 ed., 2002.
Kapitel 4: http://www.sam.math.ethz.ch/~hiptmair/tmp/Literatur1.pdf
Kapitel 6: http://www.sam.math.ethz.ch/~hiptmair/tmp/Literatur2.pdf
(In German; for English version see → Link)

Encyclopedic presentation of classic numerical integrators:

E. HAIRER, S. NORSETT, AND G. WANNER, *Solving Ordinary Differential Equations I. Nonstiff Problems*, Springer, Heidelberg, 2nd ed., 1993.

E. HAIRER AND G. WANNER, *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*, Springer, Heidelberg, 1991.

Comprehensive description of “structure preserving” integrators:

E. HAIRER, C. LUBICH, AND G. WANNER, *Geometric numerical integration*, vol. 31 of Springer Series in Computational Mathematics, Springer, Heidelberg, 2002.

Good introduction to the numerics of Hamilton's differential equations:

B. LEIMKUHLER AND S. REICH, *Simulating Hamiltonian Dynamics*, vol. 14 of Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, UK, 2004.

Advanced courses in numerics

[Spring Semester 2016]

- Prof. R. Abgrall: **Numerical Methods for Hyperbolic Partial Differential Equations**

  Core Course BSc Math, 401-3652-00L, 10 credits, 4V + 1U
- Prof. H. Ammari: **Mathematics of Super-Resolution Biomedical Imaging**

  Elective Course MSc Math, 401-4788-16L, 8 credits, 4V

If you find mistakes in the lecture notes

E-mail to Wei Wu `wei.wu@sam.math.ethz.ch` or
Zeljko Kereta `zeljko.kereta@sam.math.ethz.ch`
Subject: NUMODE Error

0.1 Acknowledgement

We thank Ms. Evgenia Ageeva for her work on the MATLAB-codes in the numerical examples. Ms. J. Mitrovic has contributed with additional explanations in the correspondingly marked boxes.

1 Introduction

We assume basic knowledge of the theory of initial value problems for ordinary differential equations for this course. These basics are taught in the courses analysis I & II in the first year of the bachelor studies in Mathematics. They are contained in [5, Kap. 8 & Sekt. 11.6], which we recommend for repetition.

In the first chapter of this course, we review the theory of initial value problems for ordinary differential equations and present important examples. We discuss the behavior of simple numerical methods using these examples.

P. Grohs

rev 63606, February 17, 2014

1.1 Initial value problems (IVP)

A first-order ordinary differential equation (ODE) is given by:

$$\boxed{\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})} \tag{1.1.1}$$

In this course, we use the following terminology, compare [12]:

- $\mathbf{f} : I \times D \mapsto \mathbb{R}^d \hat{=}$ right hand side $\quad (d \in \mathbb{N})$
- $I \subset \mathbb{R} \hat{=}$ (time)interval $\quad \leftrightarrow \quad$ "time variable" t
- open subset $D \subset \mathbb{R}^d \hat{=}$ state space/phase space
  $\leftrightarrow \quad$ "state variable" $\mathbf{y}$ (describes the "state" of a system by means of d real numbers)
- $\Omega := I \times D \hat{=}$ augmented state space $\quad$ (contains tuples $(t, \mathbf{y})$)

P. Grohs

rev 63606, February 17, 2014

✎ Notation (Newton): point $\dot{}$ $\hat{=}$ (total) derivative with respect to time t

✎ Notation: bold printing for column vectors (components are selected by subscripts, e.g., $\mathbf{y} = (y_1, \dots, y_d)^T \in \mathbb{R}^d$)

- For $d = 1$, (1.1.1) is a scalar ordinary differential equation.
- For $d > 1$, (1.1.1) is also referred to as system of ordinary differential equations:

$$(1.1.1) \quad \Longleftrightarrow \quad \frac{d}{dt}\begin{pmatrix} y_1 \\ \vdots \\ y_d \end{pmatrix} = \begin{pmatrix} f_1(t, y_1, \dots, y_d) \\ \vdots \\ f_d(t, y_1, \dots, y_d) \end{pmatrix} .$$

General assumption: continuous right hand side $\mathbf{f} : I \times D \mapsto \mathbb{R}^d$

Definition 1.1.2 (Solution of an ordinary differential equation)**.**
A function $\mathbf{y} \in C^1(J, D)$, $J \subset I$ *interval of positive length, is called* solution *of the ordinary differential equation* (1.1.1)*, if*

$$\dot{\mathbf{y}}(t) = \mathbf{f}(t, \mathbf{y}(t)) \quad \textit{for all} \quad t \in J \ .$$

[Beitrag von J. Mitrovic, FS 2011]

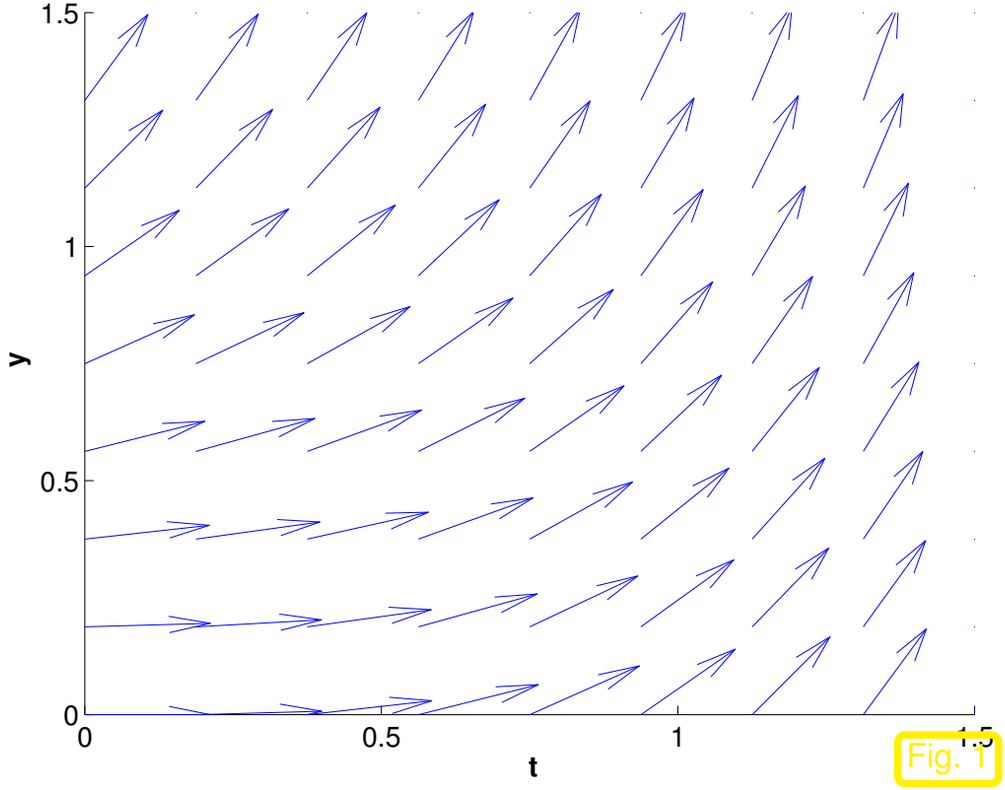
Remark: 1) $J \subseteq I$ can also be a strict subset, i.e., we call $\mathbf{y}$ a solution of the ODE, even though it fulfills the ODE only on a part of the given time interval.
2) Implicitly, the definition of solution assumes that $\mathbf{y}$ is differentiable with respect to time (at least on J).

Example 1.1.3 (Slope field and solution curves)*.*

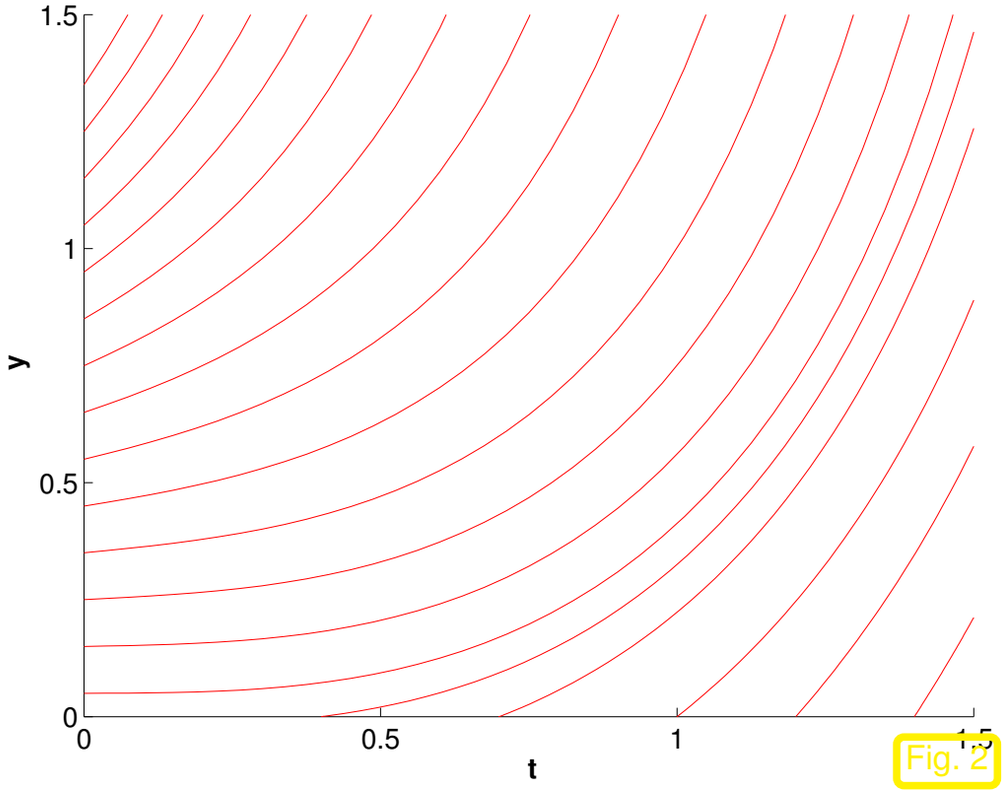
Riccati-differential equation

scalar ODE

$$\dot{y} = y^2 + t^2 \quad \blacktriangleright \quad d = 1, \quad I, D = \mathbb{R}^+ \ . \tag{1.1.4}$$



Slope field, [5, Fig. 8.1.4]



Solution curves

P. Grohs

rev 63606, February 17, 2014

Solution curves are tangential to the slope field in every point of the augmented state space.

Alternatively, one can interpret the *slope field as velocity field* of a liquid: the solution curves are the trajectories of the particles, which are advected by the liquid.

Listing 1.1: Generation of the graphics for example 1.1.3

```
% MATLAB function plotting the vector field and the solution curves for the
```

```
% Ricatti differential equation for Example 1.1.3
function Ricatti

% define the right hand side of the differential equation as function handle
fn = @(t,x) x.^2+t^2;

% plot solution curves
figure('Name','Ricatti'); hold on;
% run ode45 and plot results for different starting values on y-axis
for v = 0.05:0.1:1.4
  [t,y] = ode45(fn,[0 1.5],v);
  plot(t,y,'r-');
end
% run ode45 and plot resuts for different starting values on x-axis
for v = [0.4 0.7 1.0 1.2 1.4]
  [t,y] = ode45(fn,[v 1.5],0);
  plot(t,y,'r-');
end
% set axes, labels, ...
set(gca,'fontsize',14); axis([0 1.5 0 1.5]);
xlabel('{\bf t}'); ylabel('{\bf y}');
% Create EPS output file
print -depsc2 'riccatti1.eps'
```

```
% plot tangent field
figure('Name','LV field'); hold on;
% set up the sampling grid
N = 8; [X,Y] = meshgrid(0:1.5/N:1.5,0:1.5/N:1.5);
U = zeros(size(X)); V = zeros(size(Y));
% get velocity vectors
for i=0:N-1
  for j=0:N-1
    x = [1;fn(X(i+1,j+1),Y(i+1,j+1))];
    x = 0.3*x/norm(x);
    U(i+1,j+1) = x(1); V(i+1,j+1) = x(2);
  end
end

% plot velocity vectors
quiver(X,Y,U,V,'b-');
% set axes, labels, ...
set(gca,'fontsize',14); axis([0 1.5 0 1.5]);
xlabel('{\bf t}'); ylabel('{\bf y}');
% Create EPS output file
print -depsc2 'riccatti2.eps'
```

[Beitrag von J. Mitrovic, FS 2011]

One draws the slope field by assigning a vector $f(t,y)$ to every point (t,y). The slope field describes the qualitative behavior of the solution.

Special case: $\mathbf{f}(t,\mathbf{y}) = \mathbf{f}(\mathbf{y})$ ➡ autonomous differential equation (here $I = \mathbb{R}$)

$$\boxed{\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})} \quad . \tag{1.1.5}$$

Here: $\mathbf{f} : D \subset \mathbb{R}^d \mapsto \mathbb{R}^d$ is a (continuous) vector field ("velocity field", compare Example 1.1.3).

Remark 1.1.6 (Translation invariance of solutions of autonomous differential eqns).

$$t \mapsto \mathbf{y}(t) \text{ solution of (1.1.5) on } J \quad \Rightarrow \quad t \mapsto \mathbf{y}(t+\tau) \text{ solution (1.1.5) on } J - \tau \quad \forall \tau \in \mathbb{R}$$

[Beitrag von J. Mitrovic, FS 2011]

Due to the translation invariance, we can generate arbitrarily many solutions out of a given solution by shifting by a $t \in \mathbb{R}$.

△

Remark 1.1.7 (Rewrite as autonomous ODE).

$$\mathbf{z}(t) := \begin{pmatrix} \mathbf{y}(t) \\ t \end{pmatrix} = \begin{pmatrix} \mathbf{z}' \\ z_{d+1} \end{pmatrix} : \quad (1.1.1) \quad \leftrightarrow \quad \dot{\mathbf{z}} = \mathbf{g}(\mathbf{z}) \ , \quad \mathbf{g}(\mathbf{z}) := \begin{pmatrix} \mathbf{f}(z_{d+1}, \mathbf{z}') \\ 1 \end{pmatrix} \ . \tag{1.1.8}$$

Here, $\mathbf{z}'$ denotes the first d components of the column vector $\mathbf{z}$.

Therefore, if $\mathbf{y} \in C^1(J, D)$ is a solution of (1.1.1), then $\mathbf{z} \in C^1(J, D \times J)$ is a solution of $\dot{\mathbf{z}} = \mathbf{g}(\mathbf{z})$.

△

Generalization: A nth order ordinary differential equation , $n \in \mathbb{N}$:

$$\boxed{\mathbf{y}^{(n)} = \mathbf{f}(t, \mathbf{y}, \dot{\mathbf{y}}, \dots, \mathbf{y}^{(n-1)})} \quad . \tag{1.1.9}$$

✎ Notation: Superscript $^{(n)} \hat{=} n^{\text{th}}$ derivative with respect to time t

➤ Transformation to first-order ODE (system!) ($d \leftarrow n \cdot d$):

$$\mathbf{z}(t) := \begin{pmatrix} \mathbf{y}(t) \\ \mathbf{y}^{(1)}(t) \\ \vdots \\ \mathbf{y}^{(n-1)}(t) \end{pmatrix} = \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \vdots \\ \mathbf{z}_n \end{pmatrix} \in \mathbb{R}^{dn}\text{: (1.1.9)} \quad \leftrightarrow \quad \dot{\mathbf{z}} = \mathbf{g}(\mathbf{z}) \ , \quad \mathbf{g}(t, \mathbf{z}) := \begin{pmatrix} \mathbf{z}_2 \\ \mathbf{z}_3 \\ \vdots \\ \mathbf{z}_n \\ \mathbf{f}(t, \mathbf{z}_1, \dots, \mathbf{z}_n) \end{pmatrix} . \tag{1.1.10}$$

theory numerics	for first-order ODEs	➥	theory numerics	for n^{th} order ODEs !

Caution: (1.1.10) has a *special structure*, which a standard method for first-order ODEs might not be able to exploit in order to improve the accuracy/reduce the work load ($\rightarrow$ discussion in subsequent chapters).

Remark 1.1.11. The transformation (1.1.10) is only one out of many possibilities for transforming (1.1.9) to a first-order ODE.

△

Analysis: *symbolic computation* (separation of variables, variation of constants) generates a general solution of an ODE as parameter dependent family of curves. To give an example, one obtains formally (using chain rule) for a scalar autonomous ODE the indefinite integral

$$\dot{y} = f(y) \quad \Rightarrow \quad \frac{d}{dt}G(y) = 1 \quad \Rightarrow \quad G(y) = t + C \quad \Rightarrow \quad y(t) = G^{-1}(t + C) \;, \tag{1.1.12}$$

$$\text{with} \qquad G(\eta) = \int_{\eta_0}^{\eta} \frac{1}{f(\xi)} \, \mathrm{d}\xi \;,$$

where we assume $f(y) \neq 0$.

The issue is that a symbolic representation of G or G^{-1} is typically not available.

Therefore, we depend on *numerical solutions* of the ODE. Such a solution, however, can only be an approximation to the specific function. Consequently, numerical considerations are constrained to well-posed problems(, for which existence and uniqueness of solutions can be guaranteed). This is the case if the ODE is complemented by initial values.

ODE **+** initial values **=** Initial value problem (IVP)

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}) \quad , \quad \mathbf{y}(t_0) = \mathbf{y}_0 \quad \text{for a} \quad (t_0, \mathbf{y}_0) \in \Omega \; . \tag{1.1.13}$$

Definition 1.1.14 (Solution of an initial value problem)**.**
A solution $\mathbf{y} : J \mapsto D$, $t_0 \in J$, *of* (1.1.1)*, which fulfills* $\mathbf{y}(t_0) = \mathbf{y}_0$*, is called* *solution* *of the initial value problem* (1.1.13)*.*

Remark 1.1.15. If the ODE is autonomous $\overset{\text{Rem. 1.1.6}}{\succ}$ we use WLOG $t_0 = 0$ in (1.1.13). △

Remark 1.1.16 (Initial values for higher-order diff. eqns.).

Initial value problems for n^{th} order ordinary differential equations (1.1.9):

$$\mathbf{y}^{(n)} = f(t, \mathbf{y}) \quad , \quad \mathbf{y}(t_0) = \mathbf{y}_0 \; , \quad \dot{\mathbf{y}}(t_0) = \mathbf{y}_1, \dots, \mathbf{y}^{(n-1)}(t_0) = \mathbf{y}_{n-1} \; .$$

➡ n independent initial values need to be given.

△

1.2 Examples and basic concepts

Modeling:	Initial value problems (1.1.13) describe deterministic evolutions

1.2.1 Ecology

Example 1.2.1 (Resource-constrained growth). [2, Sect. 1.1]

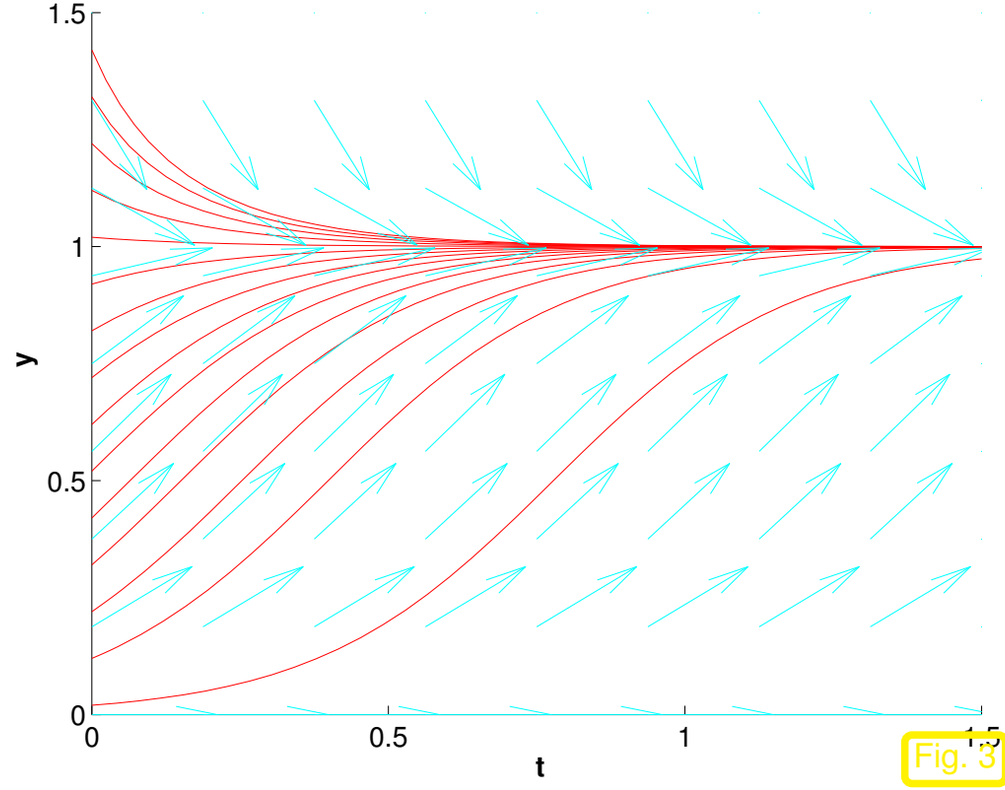
Autonomous logistic differential equation: ($d = 1$, $D = \mathbb{R}^+$, $I = \mathbb{R}$)

$$\dot{y} = (\alpha - \beta y)\, y \tag{1.2.2}$$

- $y \hat{=}$ population density, $[y] = \frac{1}{\mathrm{m}^2}$
- growth rate $\alpha - \beta y$ with growth coefficient $\alpha, \beta > 0$, $[\alpha] = \frac{1}{\mathrm{s}}$, $[\beta] = \frac{\mathrm{m}^2}{\mathrm{s}}$

In general for ODE (1.1.1):

$\mathbf{y}^* \in D$, $\mathbf{f}(t, \mathbf{y}^*) = 0 \quad \forall t \quad \succ \quad \mathbf{y}^*$ is fixed-point / stationary point for the ODE $\quad \rightarrow$ Sect. 3.2.



Slope field and solution curves ($\alpha, \beta = 5$)

- Attractive fixed-point $y = \alpha/\beta$
- Repulsive fixed-point $y = 0$

Separation of variables (1.1.12)

➥ solution of IVP for (1.2.2)

with $y(0) = y_0 > 0$

$$y(t) = \frac{\alpha y_0}{\beta y_0 + (\alpha - \beta y_0)\exp(-\alpha t)}\,, \qquad (1.2.3)$$

for all $t \in \mathbb{R}$

MATLAB-CODE: ODE-Integration

```
fn = @(t,y) 5*y*(1-y);
[t,y] = ode45(fn,[0 1.5],y0);
plot(t,y,'r-');
```

Numerical integration of the logistic differential equation in MATLAB by means of function `ode45()`:

- function handle for right hand side
- time interval $[t_0, T]$
- initial value $\mathbf{y}_0$

Return values: `t` $\hat{=}$ (column) vector of time points, `y` $\hat{=}$ (column) vector of solution values ◇

Remark 1.2.4 (IVP-solver in MATLAB). $\rightarrow$ [37]

Calling syntax:

```
[t,y] = solver(odefun,tspan,y0);
```

Function arguments:

`solver`	:	$\in$ { `ode23`, `ode45`, `ode113`, `ode15s`, `ode23s`, `ode23t`, `ode23tb` }
`odefun`	:	function handle of type `@(t,y)` $\leftrightarrow$ right hand side $\mathbf{f}(t,\mathbf{y})$
	:	(must return a *column vector*!)
`tspan`	:	2-vector $(t_0,T)^T$: initial and final time for numerical integration
`y0`	:	initial value $\mathbf{y}_0 \in \mathbb{R}^d$

Return values: `t` : (column) vector of time points $t_0 < t_1 < t_2 < \cdots < t_N = T$

`y` : $(N+1) \times d$ solution matrix, i^{th} row $\sim \mathbf{y}(t_i)$

Why does MATLAB offer that many different solvers for IVP**?**

The answer to this question and the choice of the "appropriate" solver will be a central topic of this course.

Example 1.2.5 (Predator-prey model). [2, Sect. 1.1] & [21, Sect. 1.1.1]

Autonomous Lotka-Volterra diff. eqns.: ($d = 2$)

$$\begin{aligned} \dot{u} &= (\alpha - \beta v)u \\ \dot{v} &= (\delta u - \gamma)v \end{aligned} \quad , \quad I = \mathbb{R}\,, \quad D = (\mathbb{R}^+)^2\,, \quad \alpha, \beta, \gamma, \delta > 0\,. \tag{1.2.6}$$

Population density:

$u \to$ prey,

$v \to$ predator

---

vector field $\mathbf{f}$ for Lotka-Volterra diff. eqns. ▷

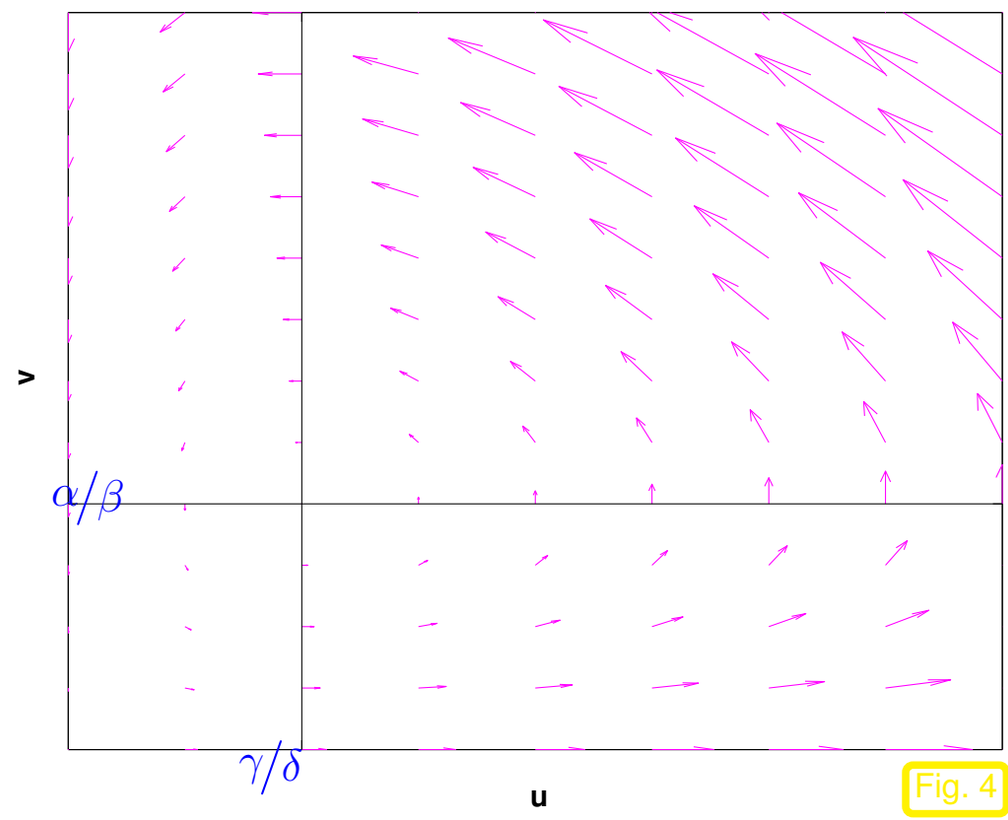
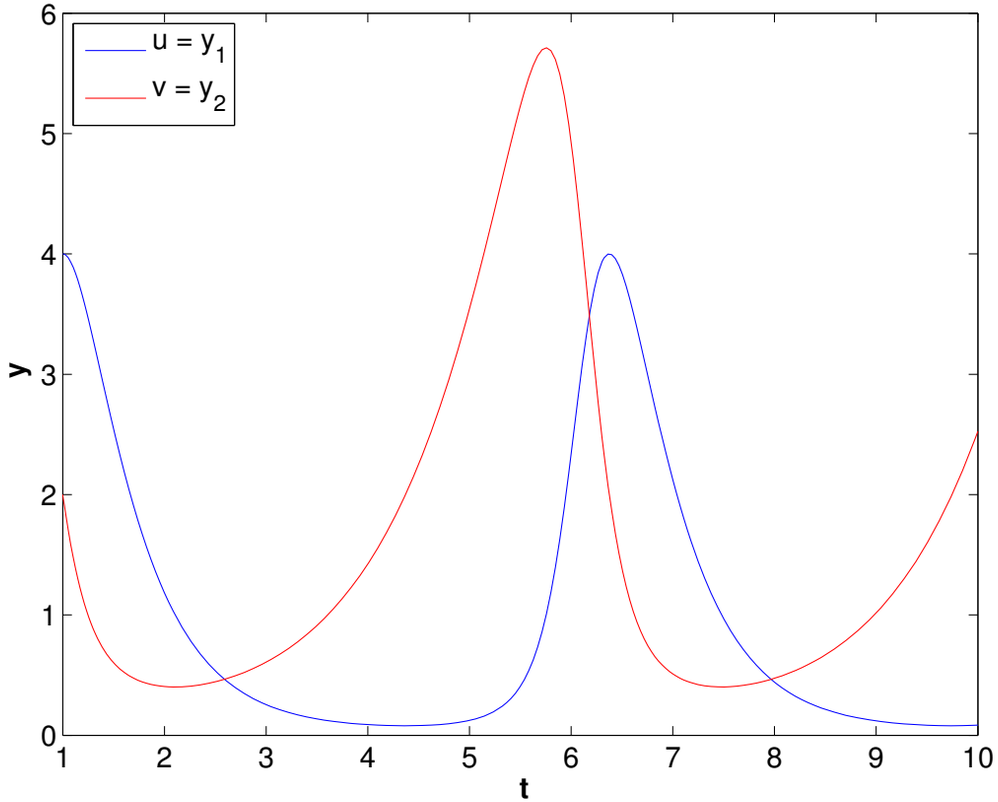


Fig. 4

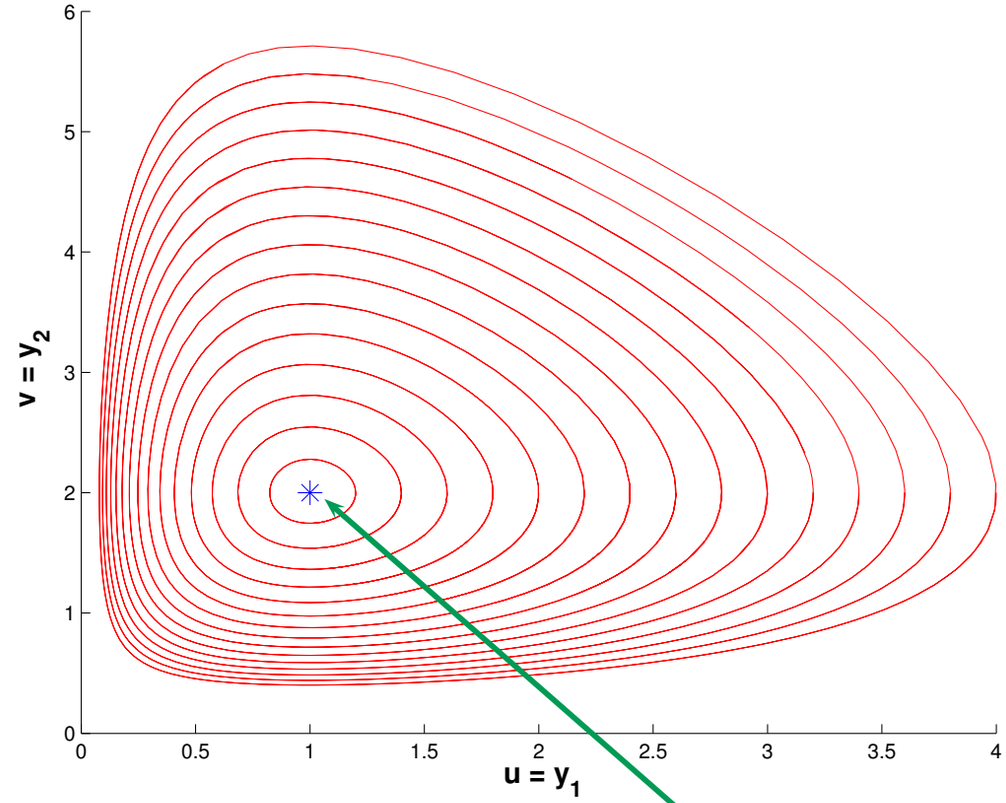
Solution curves are trajectories of particles, which are advected by the velocity field f.

$$(1.2.6) \quad \Rightarrow \quad 0 = (\delta - \frac{\gamma}{u})\dot{u} - (\frac{\alpha}{v} - \beta)\dot{v} = \frac{d}{dt}\underbrace{(\delta u - \gamma \log u - \alpha \log v + \beta v)}_{=:I(u,v)} = 0 \ .$$

- If $(u(t), v(t))$ solution of (1.2.6) $\Rightarrow$ $I(u(t), v(t)) \equiv \mathrm{const}$
- Solutions of (1.2.6) are level lines of I



Time dependent solution, $\mathbf{y}_0 := \binom{u(0)}{v(0)} = \binom{4}{2}$



solution curves of (1.2.6) $\leftrightarrow$ level lines of I

fixed-point

Closed solution curves $\leftrightarrow$ (1.2.6) has exclusively periodic solutions (for $u(0), v(0) > 0$)

$\diamond$

P. Grohs

rev 63606, February 17, 2014

Definition 1.2.7 (Invariant)**.**
A functional $I : D \mapsto \mathbb{R}$ *is called invariant of the ODE* (1.1.1)*, if*

$$I(\mathbf{y}(t)) \equiv \mathrm{const}$$

for every solution $\mathbf{y} = \mathbf{y}(t)$ *of* (1.1.1)*.*

[Beitrag von J. Mitrovic, FS 2011]

Caution: The invariant is constant for considering every solution of the ODE separately, but it is not constant for considering all solutions at the same time.

P. Grohs

rev 63606, February 17, 2014

Necessary and sufficient condition for *differentiable* invariant

$$I \text{ invariant of (1.1.1)} \quad \Leftrightarrow \quad \mathbf{grad}\, I(\mathbf{y}) \cdot \mathbf{f}(t, \mathbf{y}) = 0 \quad \forall (t, \mathbf{y}) \in \Omega \,. \tag{1.2.8}$$

Euclidean scalar product

[Beitrag von J. Mitrovic, FS 2011]

Proof: $\mathbf{grad}(I(\mathbf{y})) \cdot f(t,\mathbf{y}) = \frac{d}{dy} I(\mathbf{y}) \cdot \dot{\mathbf{y}}(t) \underbrace{=}_{\text{chain rule}} \frac{d}{dt} I(\mathbf{y}) = \frac{d}{dt} I(\mathbf{y}(t)) = 0.$

1.2.2 Chemical reaction kinetics [12, Sect. 1.3]

Example 1.2.9 (Bimolecular reaction).

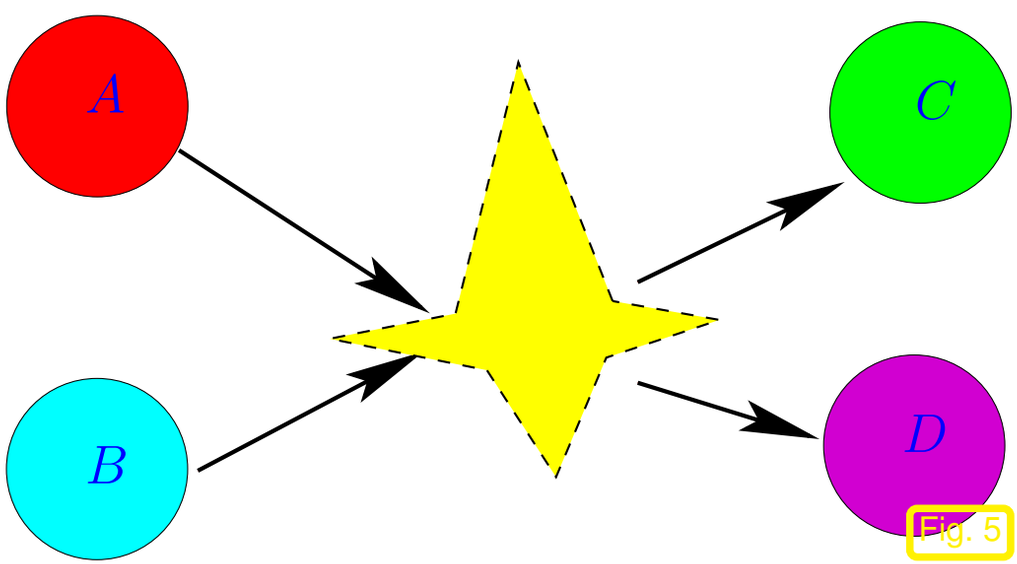


Fig. 5

Reaction: $$\mathrm{A} + \mathrm{B} \underset{k_1}{\overset{k_2}{\rightleftharpoons}} \mathrm{C} + \mathrm{D}\,. \qquad (1.2.10)$$

with reaction rates k_1 ("forward reaction"), k_2 ("reverse reaction"), $[k_1] = [k_2] = \frac{\mathrm{cm}^3}{\mathrm{mol\,s}}$.

Rule of thumb: the velocity of a bimolecular reaction is proportional to the product of concentrations of the reactants:

▶ for (1.2.10): $$\dot{c}_A = \dot{c}_B = -\dot{c}_C = -\dot{c}_D = -k_1 c_A c_B + k_2 c_C c_D \,. \tag{1.2.11}$$

$c_A, c_B, c_C, c_D \hat{=}$ (time dependent) concentrations of reactants, $[c_X] = \frac{\mathrm{mol}}{\mathrm{cm}^3}$ ➥ $c_X(t) > 0; \forall t$

(1.2.11) = autonomous ordinary diff. eqn. (1.1.5) with

$$\mathbf{y}(t) = \begin{pmatrix} c_A(t) \\ c_B(t) \\ c_C(t) \\ c_D(t) \end{pmatrix} \quad , \quad \mathbf{f}(t,\mathbf{y}) = (-k_1 y_1 y_2 + k_2 y_3 y_4) \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} \quad D := (\mathbb{R}^+)^4 \,.$$

▶ Mass conservation: $$\frac{d}{dt}\left(c_A(t) + c_B(t) + c_C(t) + c_D(t)\right) = 0$$ ◇

Example 1.2.12 (Oregonator reaction).

Special case of an oscillating (in time) Zhabotinski-Belousov reaction [15]:

$$\begin{aligned} \mathrm{BrO_3^-} + \mathrm{Br^-} &\mapsto \mathrm{HBrO_2} \\ \mathrm{HBrO_2} + \mathrm{Br^-} &\mapsto \mathrm{Org} \\ \mathrm{BrO_3^-} + \mathrm{HBrO_2} &\mapsto 2\,\mathrm{HBrO_2} + \mathrm{Ce(IV)} \\ 2\,\mathrm{HBrO_2} &\mapsto \mathrm{Org} \\ \mathrm{Ce(IV)} &\mapsto \mathrm{Br^-} \end{aligned} \tag{1.2.13}$$

$$
\begin{array}{lll}
y_1 := c(\mathrm{BrO}_3^-): & \dot{y}_1 &= -k_1 y_1 y_2 - k_3 y_1 y_3 \,, \\
y_2 := c(\mathrm{Br}^-): & \dot{y}_2 &= -k_1 y_1 y_2 - k_2 y_2 y_3 + k_5 y_5 \,, \\
y_3 := c(\mathrm{HBrO}_2): & \dot{y}_3 &= k_1 y_1 y_2 - k_2 y_2 y_3 + k_3 y_1 y_3 - 2k_4 y_3^2 \,, \\
y_4 := c(\mathrm{Org}): & \dot{y}_4 &= k_2 y_2 y_3 + k_4 y_3^2 \,, \\
y_5 := c(\mathrm{Ce(IV)}): & \dot{y}_5 &= k_3 y_1 y_3 - k_5 y_5 \,,
\end{array}
\tag{1.2.14}
$$

with (dimensionless) reaction rates:

$$
k_1 = 1.34 \,, \quad k_2 = 1.6 \cdot 10^9 \,, \quad k_3 = 8.0 \cdot 10^3 \,, \quad k_4 = 4.0 \cdot 10^7 \,, \quad k_5 = 1.0 \,.
$$

Periodic chemical reaction ➽ Movie 1, Movie 2

MATLAB-simulation with initial values $y_1(0) = 0.06$, $y_2(0) = 0.33 \cdot 10^{-6}$, $y_3(0) = 0.501 \cdot 10^{-10}$, $y_4(0) = 0.03$, $y_5(0) = 0.24 \cdot 10^{-7}$:

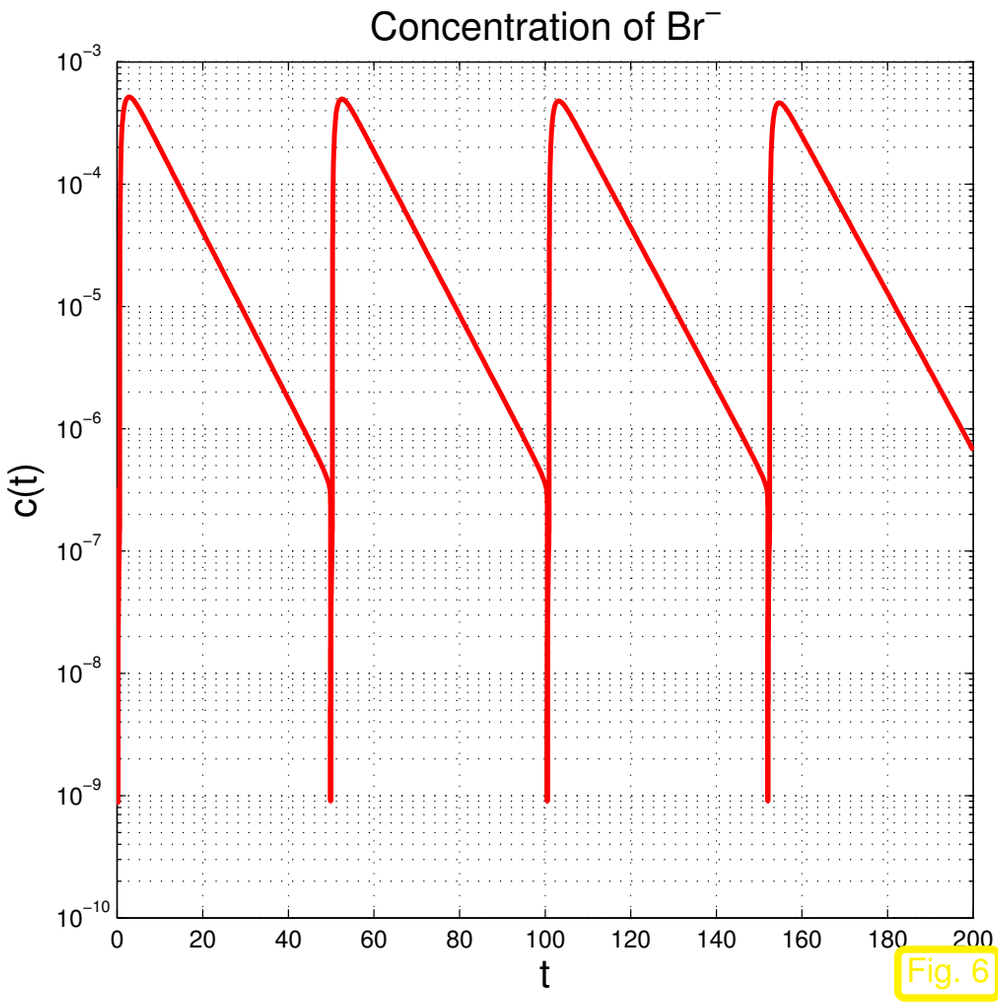


Fig. 6

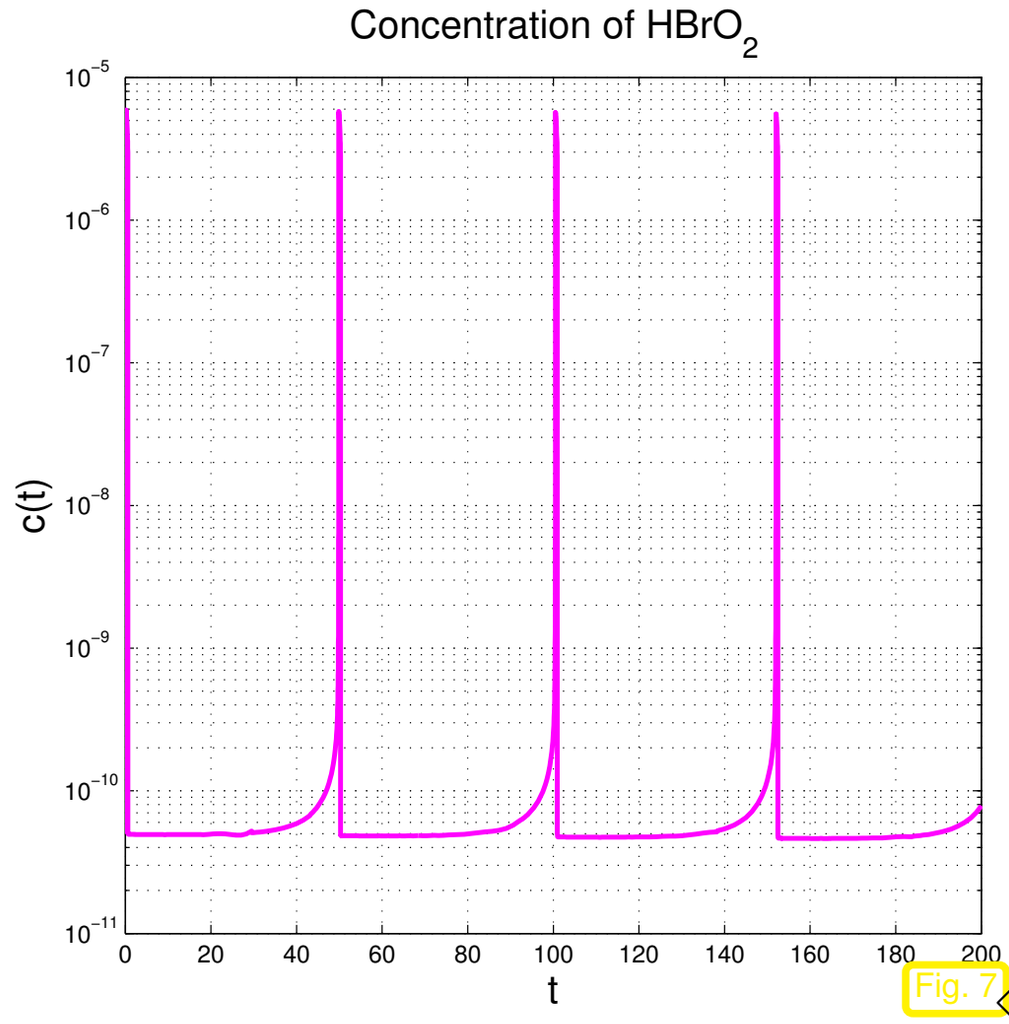


Fig. 7 ◇

P. Grohs

rev 63606, February 17, 2014

1.2.3 Physiology

Example 1.2.15 (Zeeman's model for the heartbeat). → [9, p. 655]

Quantities: $l = l(t) \;\hat{=}\;$ Length of heart muscle fiber
$p = p(t) \;\hat{=}\;$ electrochemical potential

Dimensionless phenomenological model:

$$\begin{aligned} \dot{l} &= -(l^3 - \alpha l + p)\,, \\ \dot{p} &= \beta l\,, \end{aligned} \tag{1.2.16}$$

with parameters: $\alpha \;\hat{=}\;$ tension of muscle fiber
$\beta \;\hat{=}\;$ (phenomenological) feedback parameter

Vector fields and numerical solutions for various parameters:

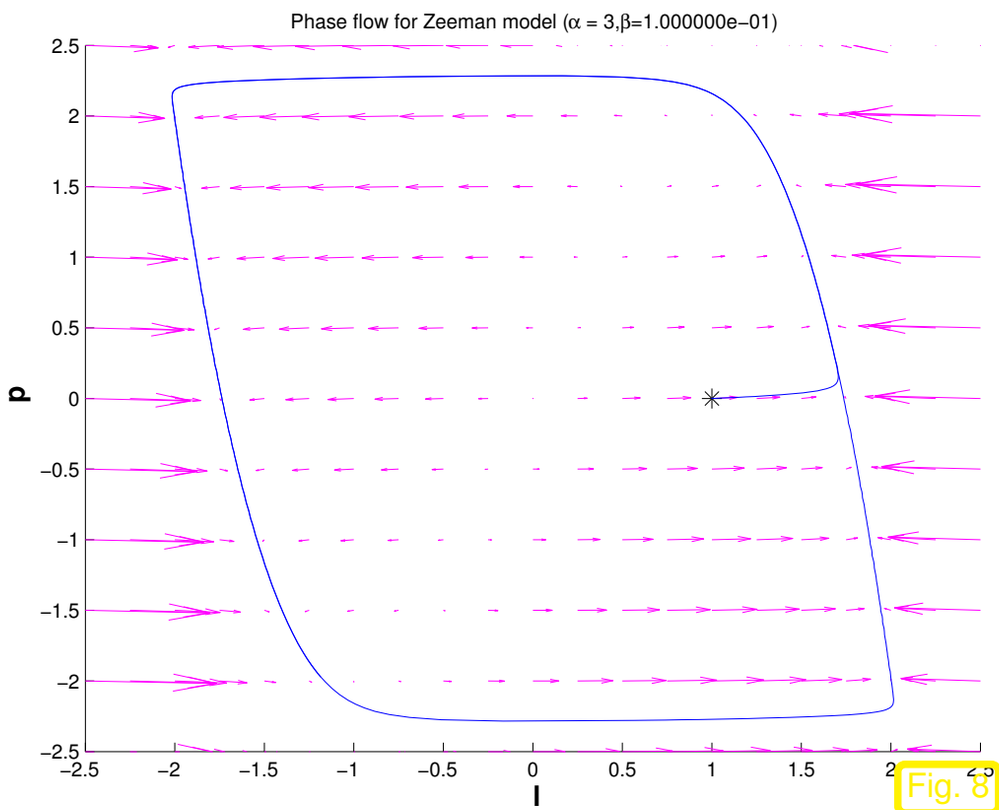


Fig. 8

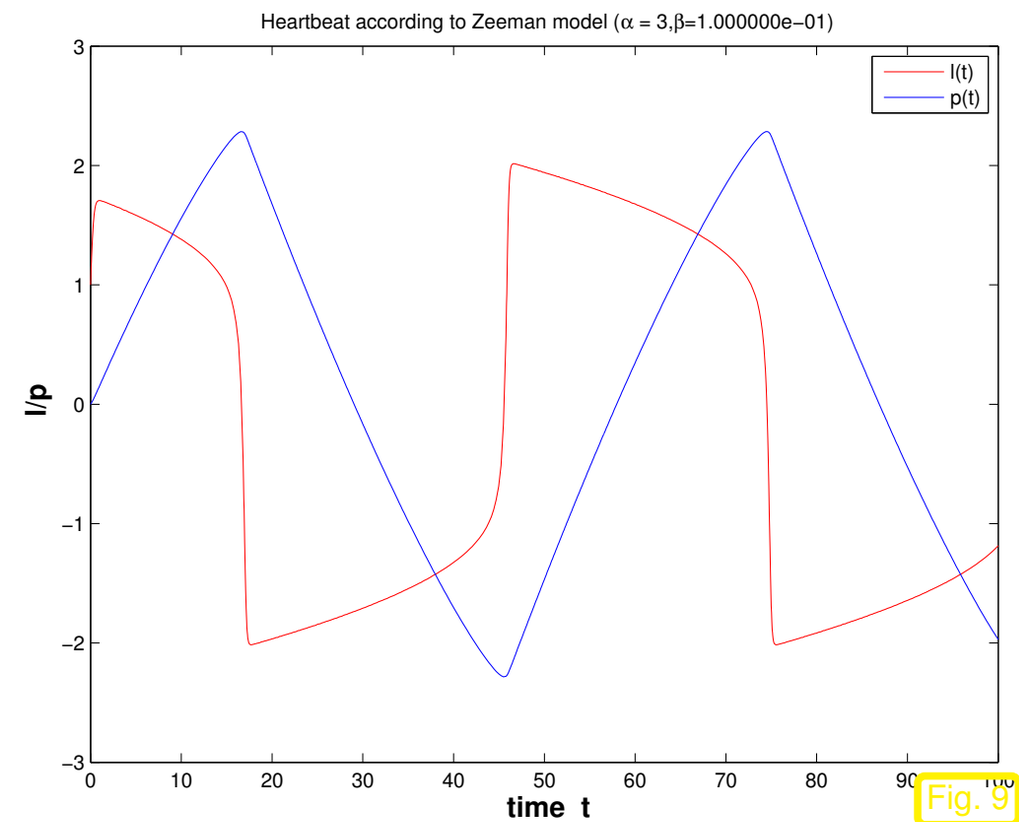


Fig. 9

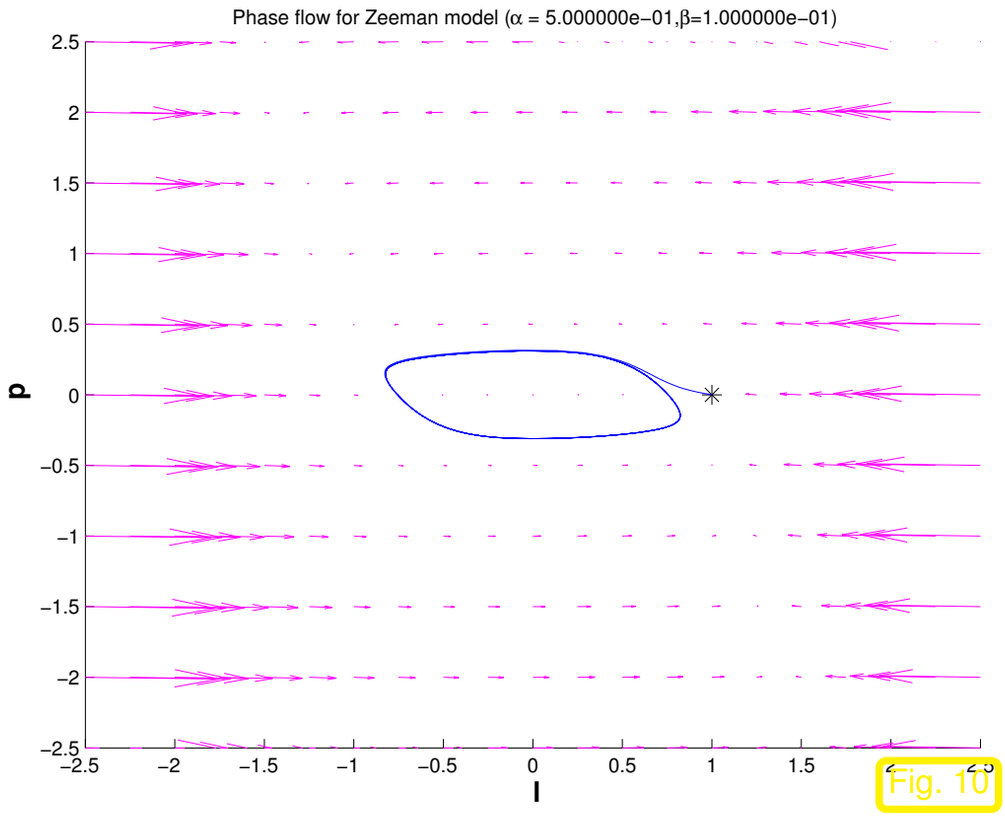


Fig. 10

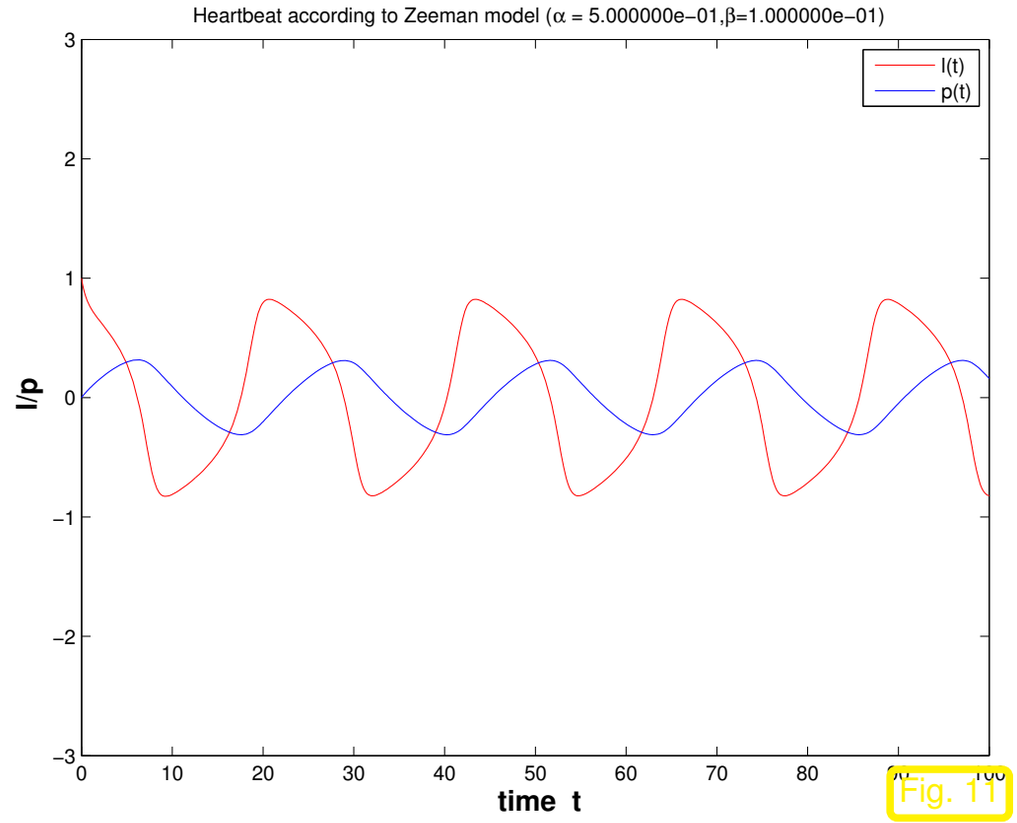


Fig. 11

Observation: $\alpha \ll 1$ ➤ "Atrial fibrillation"

Listing 1.2: Generation of the plots for example 1.2.15

```
function heartbeat
% MATLAB function for creating plots for Example 1.2.15
beat(3,'normalbeat'); beat(0.5,'crazybeat');
return
```

```
function beat(alpha,filename)
% MATLAB function for numerical simulation of the Zeeman model
% (1.2.16) of heartbeat

if (nargin < 2), filename = 'Heartbeat'; end
if (nargin < 1), alpha = 2; end

% Model equations (right hand side)

beta = 0.1; % feedback parameter
l0 = 0;     % length of relaxed muscle fibre

% Function handle to right hand side vector field
f_l = @(l,p) -(l.^3-alpha*l+p);
f_p = @(l,p) beta*(l-l0);
odefun = @(t,y) [f_l(y(1),y(2)); f_p(y(1),y(2))];

% Create plot of vector field
figure('name','heartbeat field'); hold on;
[L,P] = meshgrid((-2.5:0.25:2.5),(-2.5:0.5:2.5));
quiver(L,P,f_l(L,P),f_p(L,P),1.5,'m-');
axis([-2.5 2.5 -2.5 2.5]);
```

```
xlabel('{\bf l}','fontsize',14);
ylabel('{\bf p}','fontsize',14);
title(sprintf('Phase flow for Zeeman model (\\alpha = %d,\\beta=%d)',...
          alpha,beta));

% Compute evolution of l (length) and p (potential), see Rem. 1.2.4
tspan = [0 100]; % Duration of simulation
[t,y] = ode45(odefun,tspan,[1;0],odeset('abstol',1E-12));

% Plot trajectory of solution
plot(1,0,'k*','markersize',10);
plot(y(:,1),y(:,2),'b-');
hold off;
print('-depsc2',sprintf('%s1.eps',filename));

% Plot time-dependent solution
figure('name','heartbeat');
plot(t,y(:,1),'r-',t,y(:,2),'b-');
title(sprintf('heartbeat according to Zeeman model (\\alpha = %d,\\beta=%d)',alpha,beta));
xlabel('{\bf time  t}','fontsize',14);
ylabel('{\bf l/p}','fontsize',14);
```

```
axis([tspan -3 3]); legend('l(t)','p(t)');

print('-depsc2',sprintf('%s2.eps',filename));
```

◇

1.2.4 Mechanics

Example 1.2.17 (Mathematical pendulum). [2, I.3. Ex. (3.4c)]

State space D = configuration space for minimal coordinates (= angular displacement)

➢ $d = 1$, $D = \mathbb{T}$ (circle line = "1D Torus")

angular displacement $\alpha \in [-\pi, \pi[)$

Newton's equations of motion:

$$ml\,\ddot{\alpha}(t) = -mg\sin\alpha(t)\ . \qquad (1.2.18)$$

▶ autonomous second-order ODE, compare (1.1.9)

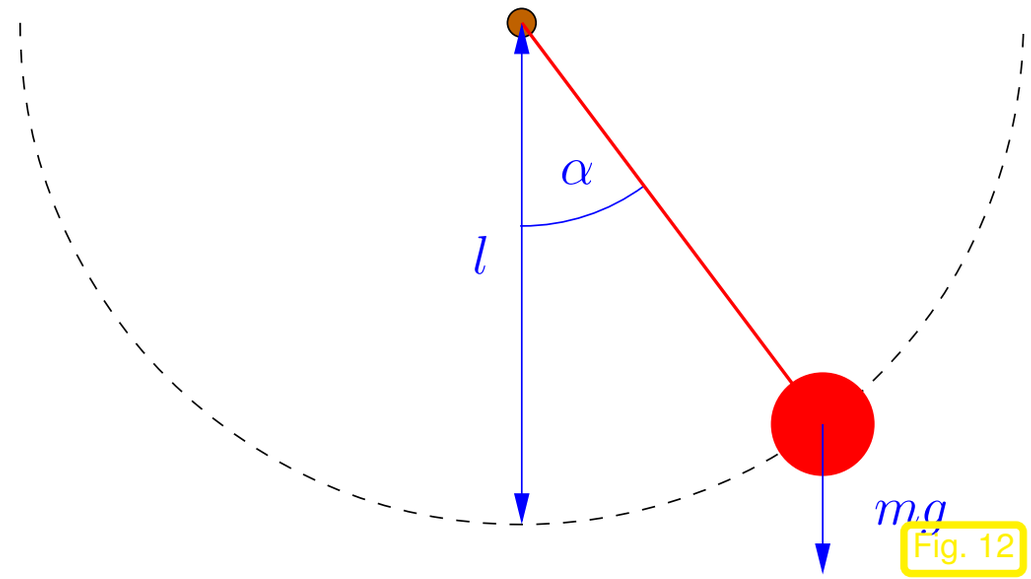


Fig. 12

Formal transformation in first-order ODE:

$$\text{angle velocity} \quad p := \dot{\alpha} \quad \Rightarrow \quad \frac{d}{dt}\begin{pmatrix}\alpha\\ p\end{pmatrix} = \begin{pmatrix} p \\ -\frac{g}{l}\sin\alpha\end{pmatrix}\ . \qquad (1.2.19)$$

Energy conservation: $E(t) = \underbrace{\tfrac{1}{2}ml^2p(t)^2}_{\text{kinetic energy } T(t)} - \underbrace{mgl\cos\alpha(t)}_{\text{potential energy } U(t)} \quad \Rightarrow \quad E(t) \equiv \text{const.}$ (invariant → Def. 1.

P. Grohs

rev 63606, February 17, 2014

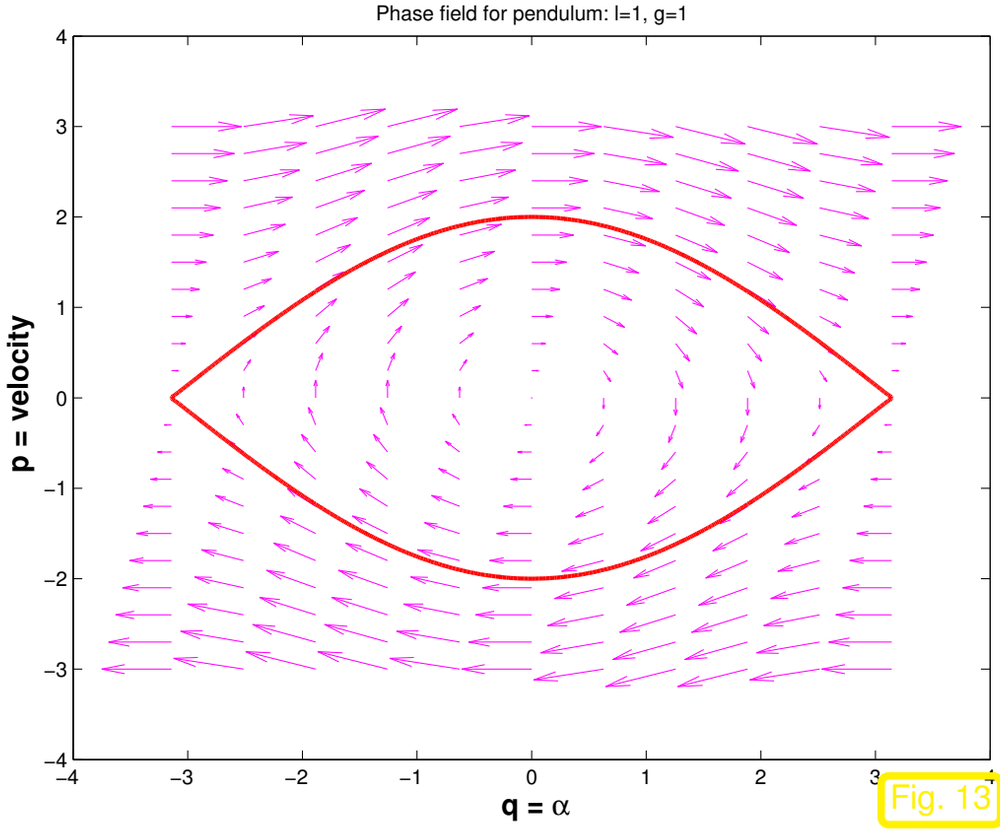


Fig. 13

vector field for (1.2.19)

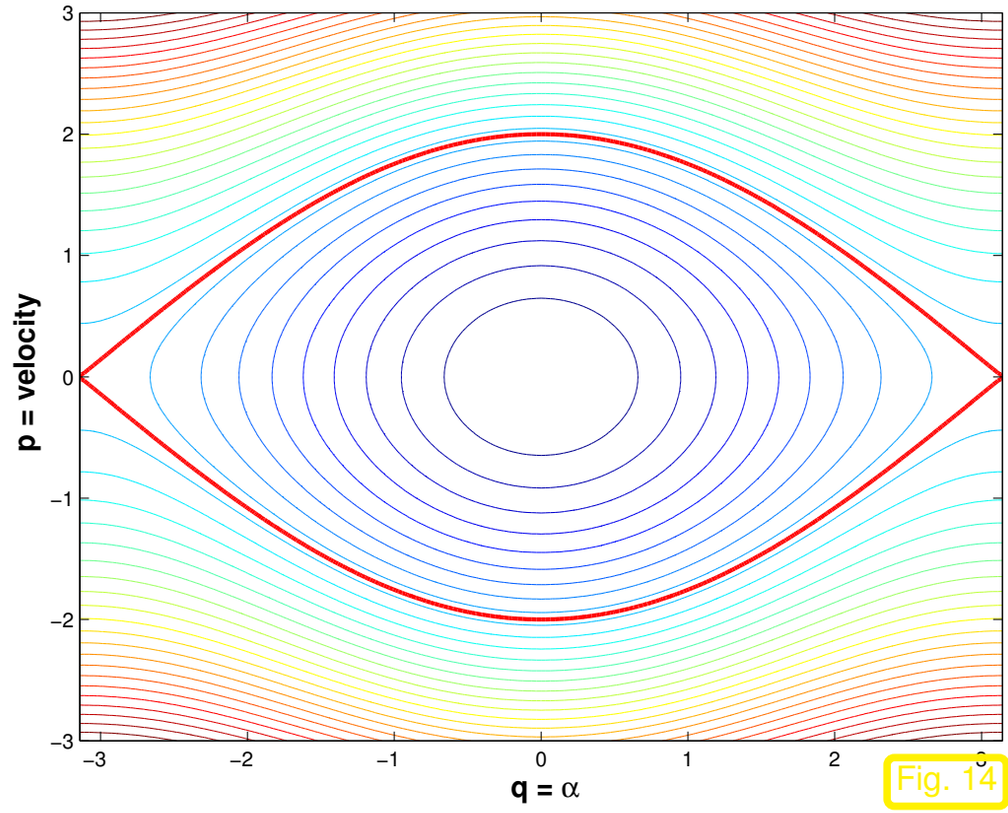


Fig. 14

Isolines of energy $\leftrightarrow$ solution curves $\diamond$

P. Grohs

rev 63606, February 17, 2014

Definition 1.2.20 (Hamilton's differential equation). $\rightarrow$ *[21, Sect. VI.1.2]*

Let $n \in \mathbb{N}$, $M \subset \mathbb{R}^n$ open, and let $H : \mathbb{R}^n \times M \mapsto \mathbb{R}$, $H = H(\mathbf{p}, \mathbf{q})$, be continuously differentiable. Then the first-order ordinary differential equation

$$\dot{\mathbf{p}}(t) = -\frac{\partial H}{\partial \mathbf{q}}(\mathbf{p}(t), \mathbf{q}(t)) \quad , \quad \dot{\mathbf{q}}(t) = \frac{\partial H}{\partial \mathbf{p}}(\mathbf{p}(t), \mathbf{q}(t)) \; , \tag{1.2.21}$$

is called an autonomous Hamiltonian system with Hamiltonian H.

Remark 1.2.22. In classical mechanics, equations of motion result in autonomous Hamiltonian systems, compare [2, Sect. I.3] for an introduction, and [3] for an extensive discussion. $\triangle$

P. Grohs

rev 63606, February 17, 2014

Lemma 1.2.23 ("Energy conservation").

The Hamiltonian H is an invariant of the autonomous Hamiltonian system.

Hamiltonian system of the form (1.1.1):

$$\mathbf{y} = \begin{pmatrix} \mathbf{p} \\ \mathbf{q} \end{pmatrix} \quad \Rightarrow \quad (1.2.21) \quad \Leftrightarrow \quad \dot{\mathbf{y}} = \mathbf{J}^{-1} \cdot \mathbf{grad}\, H(\mathbf{y}) \;, \quad \mathbf{J} := \begin{pmatrix} 0 & \mathbf{I}_n \\ -\mathbf{I}_n & 0 \end{pmatrix} \in \mathbb{R}^{2n,2n} \;. \qquad (1.2.24)$$

✎ Notation: $\mathbf{I}_n \hat{=} n \times n$ unity matrix

Together with (1.2.8), this implies Lemma 1.2.23, as $\mathbf{J}$ is *skew-symmetric* ($\mathbf{J}^T = -\mathbf{J}$) and there holds $\boldsymbol{x} \cdot \mathbf{A}\boldsymbol{x} = 0 \quad \forall \boldsymbol{x} \in \mathbb{R}^n$ for every skew-symmetric matrix $\mathbf{A} \in \mathbb{R}^{n,n}$.

[Beitrag von J. Mitrovic, FS 2011]

Alternatively, one can show this result by computation:
We use that $\mathbf{grad}(H(\mathbf{y}))^T \cdot \mathbf{J}^{-1} \cdot \mathbf{grad}(H(\mathbf{y}))$ is a number. Therefore, we can transpose it without changing the result:
$\mathbf{grad}(H(\mathbf{y}))^T \cdot \mathbf{J}^{-1} \cdot \mathbf{grad}(H(\mathbf{y})) = \mathbf{grad}(H(\mathbf{y}))^T \cdot \mathbf{J}^{-T} \cdot \mathbf{grad}(H(\mathbf{y}))$
$= \mathbf{grad}(H(\mathbf{y}))^T \cdot (-\mathbf{J}) \cdot \mathbf{grad}(H(\mathbf{y}))$. This implies that $\mathbf{grad}(H(\mathbf{y}))^T \cdot \mathbf{J}^{-1} \cdot \mathbf{grad}(H(\mathbf{y})) = 0$,or in other words that $\mathbf{grad}(H(\mathbf{y}))$ is an invariant.

Example 1.2.25 (Mass point in central field).

Newton's equations of motion of a body (space coordinate $\mathbf{r} = \mathbf{r}(t)$) with mass $m > 0$ in force field $\mathbf{f} : \mathbb{R}^n \mapsto \mathbb{R}^n$, $n \in \mathbb{N}$ is given by:

$$m\,\ddot{\mathbf{r}}(t) = \mathbf{f}(\mathbf{r}(t)) \;. \tag{1.2.26}$$

Special case: radially symmetric conservative force field

$$\mathbf{f}(\boldsymbol{x}) = -\,\mathbf{grad}\,U(\boldsymbol{x}) \;, \quad \boldsymbol{x} \in \mathbb{R}^n \quad , \quad U(\boldsymbol{x}) = G(\|\boldsymbol{x}\|) \;. \tag{1.2.27}$$

✎ Notation: $\|\boldsymbol{x}\| := \sqrt{x_1^2 + \cdots + x_n^2} \mathrel{\hat{=}}$ Euclidean norm of a vector

In particular $G(r) = -\dfrac{G_0}{r}$: Kepler's problem: [21, Sect. I.2], [12, Sect. 1.1]

movement of a planet in the gravity field of the sun

$\longleftrightarrow$ Hamiltonian system ($\rightarrow$ Def. 1.2.20) with configuration space $M := \mathbb{R}^n \setminus \{0\}$, $\mathbf{q} := \mathbf{r}$, and

Hamiltonian (energy) $$H(\mathbf{p},\mathbf{q}) := \underbrace{\tfrac{1}{2m}\,\|\mathbf{p}\|^2}_{\text{kinetic energy}} + \underbrace{G(\|\mathbf{q}\|)}_{\text{potential energy}} \tag{1.2.28}$$

$\mathbf{p} := m\dot{\mathbf{r}} \mathrel{\hat{=}}$ momentum,

P. Grohs

rev 63606, February 17, 2014

$$\dot{\mathbf{p}} = -G'(\|\mathbf{q}\|)\frac{\mathbf{q}}{\|\mathbf{q}\|} \quad , \quad \dot{\mathbf{q}} = m^{-1}\mathbf{p} \; . \tag{1.2.29}$$

Lemma 1.2.30 (Orbital plane).
Every solution $\binom{\mathbf{p}}{\mathbf{q}} : J \subset \mathbb{R} \mapsto \mathbb{R}^{2n}$ *of* (1.2.29) *fulfills*

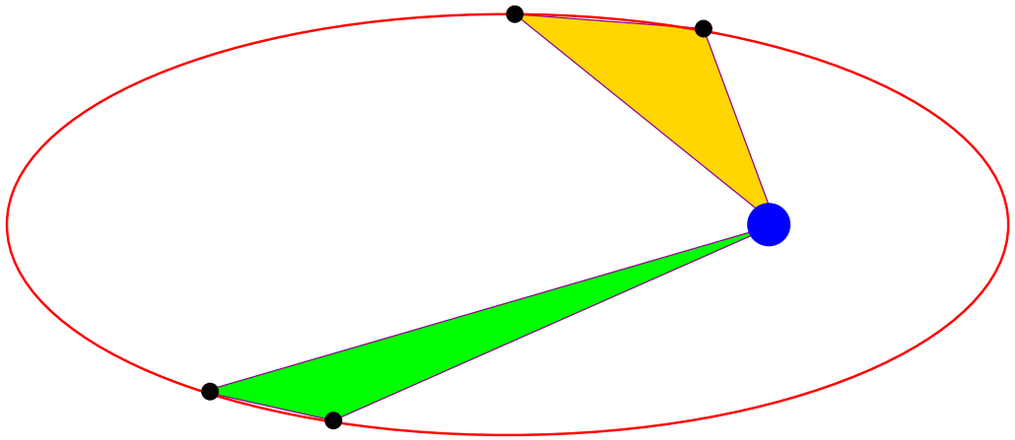
$$\mathbf{p}(t), \mathbf{q}(t) \in \operatorname{Span}\{\mathbf{p}(t_0), \mathbf{q}(t_0)\} \quad \forall t_0, t \in J \; .$$

Lemma 1.2.32 (Conservation of angular momentum).
For $n = 3$*, the angular momentum (w.r.t.* 0*)* $M := \mathbf{p} \times \mathbf{q}$ *is an invariant (*$\rightarrow$ *Def. 1.2.7) of* (1.2.29).

✎ Notation: $\times \hat{=}$ vector product in $\mathbb{R}^3$:

$$\mathbf{a} \times \mathbf{b} = \begin{pmatrix} a_2 b_3 - a_3 b_2 \\ a_3 b_1 - a_1 b_3 \\ a_1 b_2 - a_2 b_1 \end{pmatrix} \; .$$

Theorem 1.2.33 (Kepler's 2nd law). *If* $t \mapsto \mathbf{q}(t)$ *solves the differential equation* (1.2.29), *then the vector* $\mathbf{q}(t)$ *sweeps out equal areas in the orbital plane during equal intervals of time.*



Kepler's 1st law (for gravity potentials):

For $G(r) = -\frac{G_0}{r}$, the solution curves of (1.2.26) are ellipses with focus 0.

Outlook: To simulate the movement of planets in our solar system, one would need to integrate (1.2.29) for a long time span. Unfortunately, such an integration is not trivial. Even though both the implicit Euler method and the Störmer-Verlet method preserve the angular momentum exactly, this is not the case for the energy of the system (Hamiltonian) [21, Table I.2.1]. This issue will be discussed in more detail in section 4.4.

1.3 Theory [12, Sect. 2], [2, Ch. II]

For given right hand side $f : \Omega := I \times D \mapsto \mathbb{R}^d$, $d \in \mathbb{N}$, $I \subset \mathbb{R}$ open interval, $D \subset \mathbb{R}^d$ open set, consider the initial value problem

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}) \quad , \quad \mathbf{y}(t_0) = \mathbf{y}_0 \quad \text{for given} \quad (t_0, \mathbf{y}_0) \in \Omega \; . \tag{1.1.13}$$

1.3.1 Existence and uniqueness of solutions

Two important concepts:

Definition 1.3.1 (Maximal continuation of a solution)**.**
A solution $\mathbf{y} \in C^1([t_0, t_+[, D)$ *(→ Def. 1.1.2) of the IVP is called maximally continued (into the future), if exactly one of the following three cases holds true*

(i)	$t_+ = \infty$	*(solution exists for all times)*
(ii)	$t_+ < \infty$, $\lim\limits_{t \to t_+} \|\widetilde{\mathbf{y}}(t)\| = \infty$	*("blow-up")*
(iii)	$t_+ < \infty$, $\lim\limits_{t \to t_+} \operatorname{dist}((t, \widetilde{\mathbf{y}}(t)), \partial\Omega) = 0$.	*("collapse")*

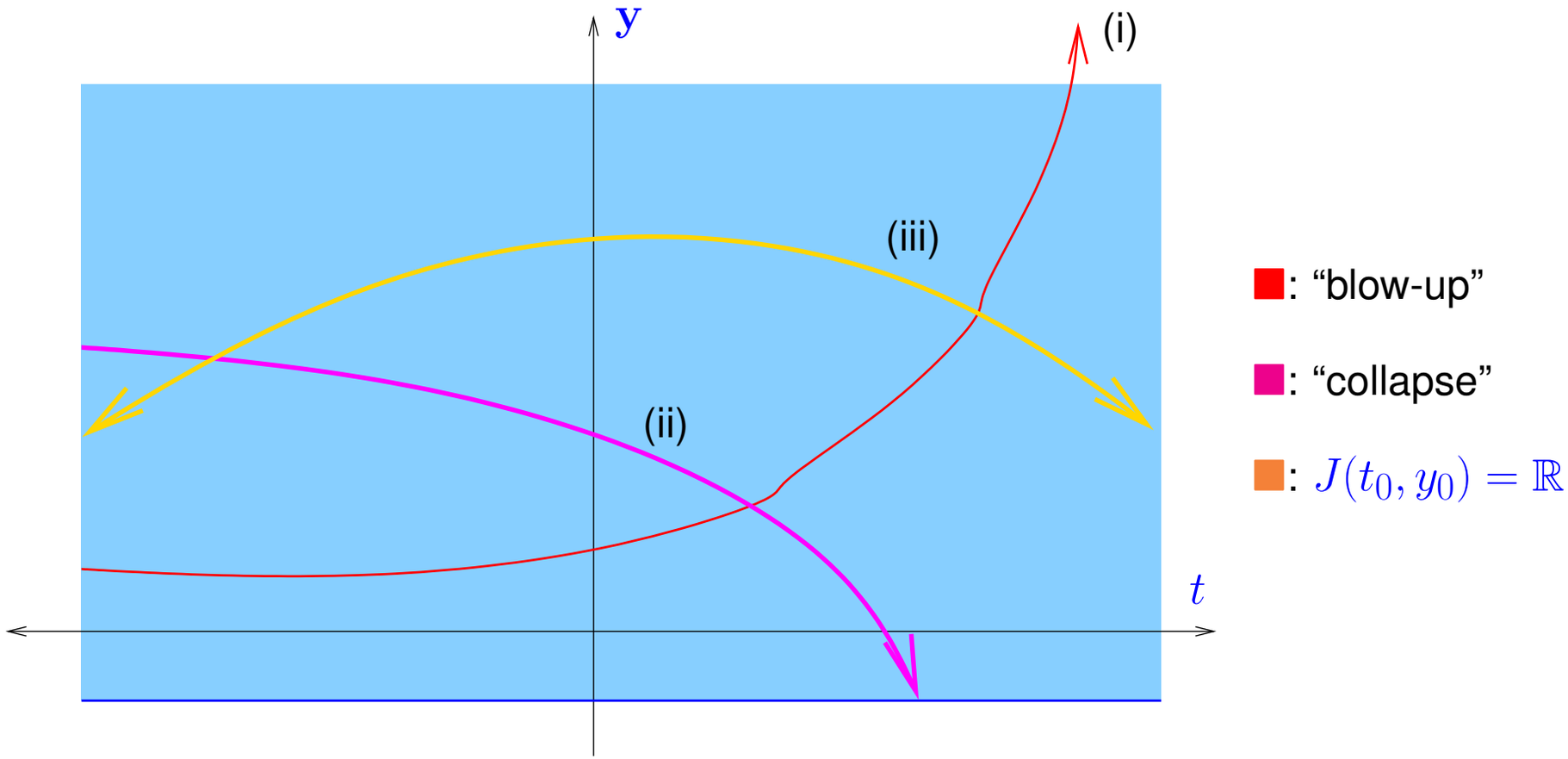
(Analogous: Maximally continued into the past $]t_-, t_0]$*)*

P. Grohs

rev 63606, February 17, 2014

Reminder: $\Omega := I \times D$ is the augmented state space

➢ collapse ↔ solution goes to the boundary of the augmented state space! (Typically to the spot for which the right hand side $\mathbf{f}$ is not defined anymore.)



Notation: $J(t_0, \mathbf{y}_0) =]t_-, t_+[$ = maximum interval of existence for solution of IVP (1.1.13).

Definition 1.3.2 (Local Lipschitz continuity)**.**
$\mathbf{f} : \Omega \mapsto \mathbb{R}^d$ *is called locally Lipschitz continuous (w.r.t. the state variable)*

$$\forall (t, \mathbf{y}) \in \Omega: \quad \exists \delta > 0,\ L > 0:$$
$$\Leftrightarrow: \qquad \|\mathbf{f}(\tau, \mathbf{z}) - \mathbf{f}(\tau, \mathbf{w})\| \leq L \|\mathbf{z} - \mathbf{w}\|$$
$$\forall \mathbf{z}, \mathbf{w} \in D: \ \|\mathbf{z} - \mathbf{y}\| < \delta,\ \|\mathbf{w} - \mathbf{y}\| < \delta, \forall \tau \in I: |t - \tau| < \delta \ .$$

Local Lipschitz continuity implies global Lipschitz continuity on every *compact* subset K of Ω:

$$\exists L = L(K) > 0: \quad \|\mathbf{f}(\tau, \mathbf{z}) - \mathbf{f}(\tau, \mathbf{w})\| \leq L \|\mathbf{z} - \mathbf{w}\| \quad \forall (\tau, \mathbf{z}), (\tau, \mathbf{w}) \in K \ .$$

✎ Notation: $D_{\mathbf{y}}\mathbf{f} \hat{=}$ derivative of $\mathbf{f}$ w.r.t. the state variables (= Jacobian matrix $\in \mathbb{R}^{d,d}$ **!**)

A useful criterion for local Lipschitz continuity (proof uses the mean value theorem):

Lemma 1.3.3 (Criterion for local Lipschitz continuity)**.**
Let both $\mathbf{f}$ and $D_{\mathbf{y}}\mathbf{f}$ be continuous on Ω. Then $\mathbf{f}$ is locally Lipschitz continuous ($\rightarrow$ Def. 1.3.2).

Theorem 1.3.4 (Theorem of Peano & Picard-Lindelöf)**.** *[2, Satz II(7.6)]*
Let $\mathbf{f} : \hat{\Omega} \mapsto \mathbb{R}^d$ be locally Lipschitz continuous w.r.t. the variable $\mathbf{y}$ ($\rightarrow$ Def. 1.3.2). Then the IVP (1.1.13) *with arbitrary initial values $(t_0, \mathbf{y}_0) \in \Omega$ has a unique, maximally extended ($\rightarrow$ Def. 1.3.1) solution $\mathbf{y} : J(t_0, \mathbf{y}_0) \mapsto D$.*

P. Grohs

rev 63606, February 17, 2014

Key idea of the proof: ($\rightarrow$ [38, I.§6], [5, Sekt. 11.6]) Integration of (1.1.13) implies

$$\mathbf{y}(t) = \mathbf{y}_0 + \int_{t_0}^{t} \mathbf{f}(s, \mathbf{y}(s))\,\mathrm{d}s, \quad t \geq t_0. \tag{1.3.5}$$

Define the space

$$\mathcal{F} = \{\mathbf{y} \in C([t_0, t_1[),\ \mathbf{y}(t_0) = \mathbf{y}_0\}$$

for a $t_1 > t_0$ and the operator

$$T : \mathcal{F} \to \mathcal{F}, \quad T : \mathbf{y} \mapsto \mathbf{z}(t) = \mathbf{y}_0 + \int_{t_0}^{t} \mathbf{f}(s, \mathbf{y}(s))\, \mathrm{d}s.$$

Then, one can rewrite (1.3.5) on the interval $[t_0, t_1]$ as fixed-point equation $T(\mathbf{y}) = \mathbf{y}$ on $\mathcal{F}$. Due to the local Lipschitz continuity, one can deduce that T is a contraction for sufficiently small $t_1 > t_0$. The Banach fixed-point theorem then implies the claim for the time interval (t_0, t_1). The maximal lifetime follows by continuation. □

Remark 1.3.6 (Domain of definition of solutions of IVPs).

" The solution of an IVP determines its domain of definition itself."

! The domain of definition $J(t_0, \mathbf{y}_0)$ typically depends on $(t_0, \mathbf{y}_0)$!

Terminology: If $J(t_0, \mathbf{y}_0) = I$ ➥ solution $\mathbf{y} : I \mapsto \mathbb{R}^d$ is global.

△

Definition 1.3.7 (Evolution).
The two-parameter family $\mathbf{\Phi}^{s,t}$ of mappings $\mathbf{\Phi}^{s,t} : D \mapsto D$ is called evolution for the differential equation $\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y})$, if

$$t \in J(s,\mathbf{z}) \mapsto \mathbf{\Phi}^{s,t}\mathbf{z} \quad \text{solution of the IVP} \quad \begin{array}{l} \dot{\mathbf{y}} = f(t,\mathbf{y}) \,, \\ \mathbf{y}(s) = \mathbf{z} \,, \end{array} \quad \text{for all} \quad (s,\mathbf{z}) \in \Omega \,.$$

Domain of definition: $$\Phi : \left\{ \begin{array}{rcl} \widetilde{\Omega} & \mapsto & D \\ (t,s,\mathbf{y}) & \mapsto & \Phi^{s,t}\mathbf{y} \end{array} \right. , \quad \widetilde{\Omega} := \bigcup_{(s,\mathbf{y})\in\Omega} J(s,\mathbf{y}) \times \{(s,\mathbf{y})\}$$

Theorem 1.3.4 $\Rightarrow$ $$\boxed{\mathbf{\Phi}^{t,t} = \mathrm{Id} \quad , \quad \mathbf{\Phi}^{s,t}\mathbf{y} = (\mathbf{\Phi}^{r,t} \circ \mathbf{\Phi}^{s,r})\mathbf{y} \,, \quad t,r \in J(s,\mathbf{y}),\, (s,\mathbf{y}) \in \Omega} \,. \tag{1.3.8}$$

Convention: For autonomous differential equations (1.1.5) ($\rightarrow$ Rem. 1.1.15): $\mathbf{\Phi}^t := \mathbf{\Phi}^{0,t}$

➢ If $J(0,\mathbf{y}) = \mathbb{R} \quad \forall \mathbf{y} \in D$, (1.3.8) implies:

Group of mappings from D: $$\mathbf{\Phi}^s \circ \mathbf{\Phi}^t = \mathbf{\Phi}^{s+t} \,, \quad \mathbf{\Phi}^{-t} \circ \mathbf{\Phi}^t = Id \quad \forall t \in \mathbb{R} \,. \tag{1.3.9}$$

Remark 1.3.10 (Numerical integrators as approximate evolutions).

MATLAB solution of an , compare Rem. 1.2.4

```
[t,y] = solver(odefun,[t0 T],y0)
```

$$\mathbf{\Phi}^{s,t}\mathbf{y}$$

☞ Numerical solution methods for initial value problems for *an* ordinary differential equation realize *approximations* of evolutions $\rightarrow$ Def. 2.1.2. △

Example 1.3.11 (Autonomous scalar differential equations). ▷ $d = 1$

- $f(t,y) = -\lambda y$, $\lambda \in \mathbb{R}$ ▶ solution of IVP $y(t) = y_0 e^{-\lambda t}$, $t \in \mathbb{R}$
  - exists for all times, i.e. $]t_-, t_+[= \mathbb{R}$ for every y_0: global solution.
  - Corresponding evolution:

$$\mathbf{\Phi}^t : \mathbb{R} \mapsto \mathbb{R} \quad , \quad \mathbf{\Phi}^t(y_0) = e^{-\lambda t} y_0 \; .$$

- $f(t,y) = \lambda y^2$, $\lambda \in \mathbb{R}$: $\dot{y} = \lambda y^2$, $y(0) = y_0 \in \mathbb{R}$

P. Grohs

rev 63606, February 17, 2014

This ODE describes self-reinforcing growth: “explosion equation”

Solution:

$$y(t) = \begin{cases} \dfrac{1}{y_0^{-1} - \lambda t} & \text{, if } y_0 \neq 0 \text{ ,} \quad \text{(blow-up)} \\ 0 & \text{, if } y_0 = 0 \text{ .} \end{cases} ,$$

$$\lambda, y_0 > 0 \quad \Rightarrow \quad J(0, y_0) =]-\infty, 1/\lambda y_0[\text{ .}$$

Solution curves for $\lambda = 1$ ▷

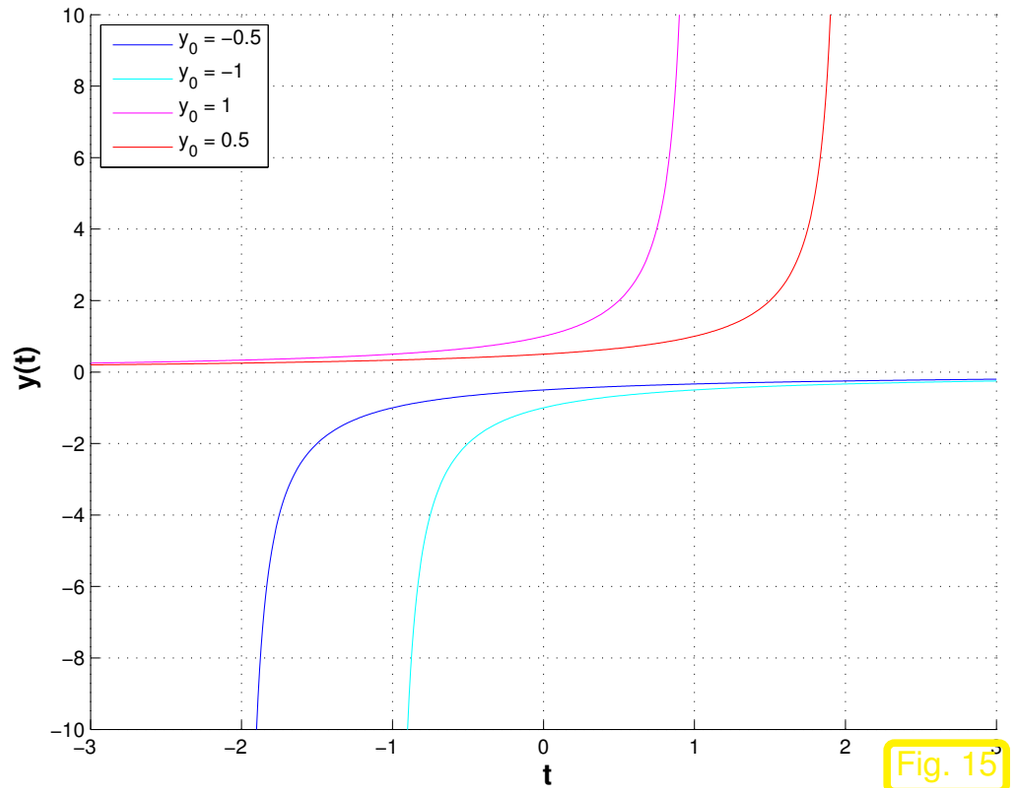


Fig. 15

- $f(t, y) = -\frac{1}{\sqrt{y}}$, $D = \mathbb{R}^+$, initial value $y(0) = 1$

  ➢ $y(t) = (1 - 3t/2)^{2/3}$, $t_- = -\infty$, $t_+ = 2/3$

  (Solution goes to the boundary $y = 0$ of the augmented state space: collapse)

- $f(t, y) = \sin(1/y) - 2$, $D = \mathbb{R}^+$, initial value $y(0) = 1$ [12, Bsp. 2.14]

  Solution $y(t)$ satisfies $\dot{y} \leq -1$ ➥ $y(t) \leq 1 - t$ ➥ collapse for $t^* < 1$.

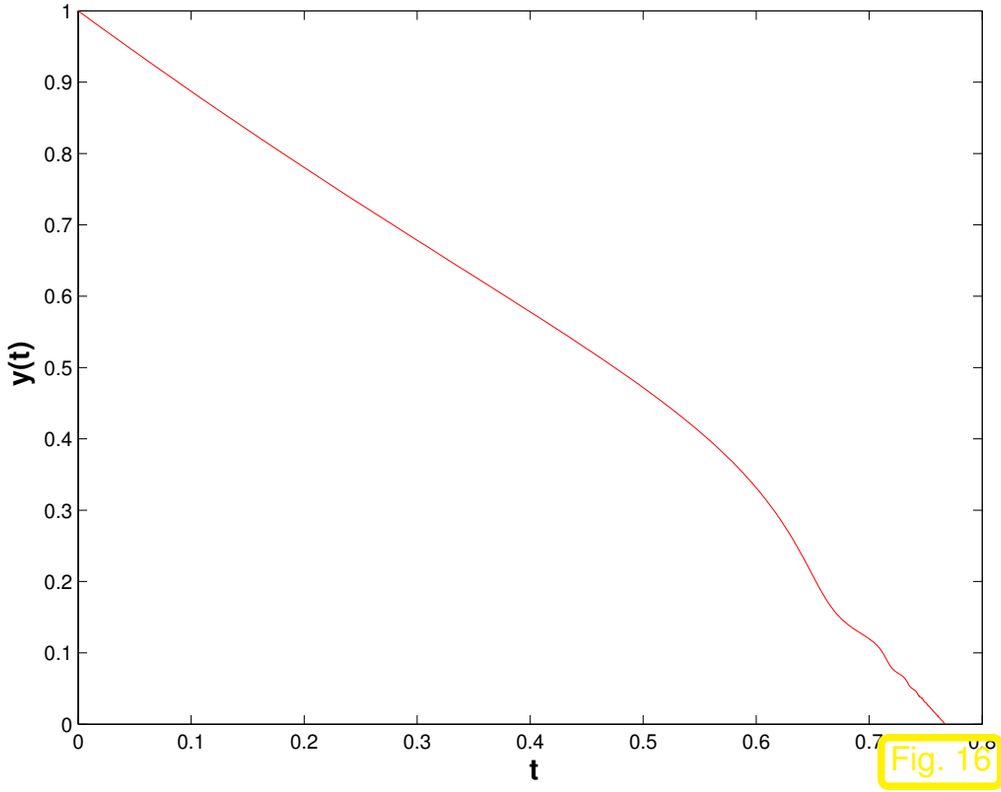


Fig. 16

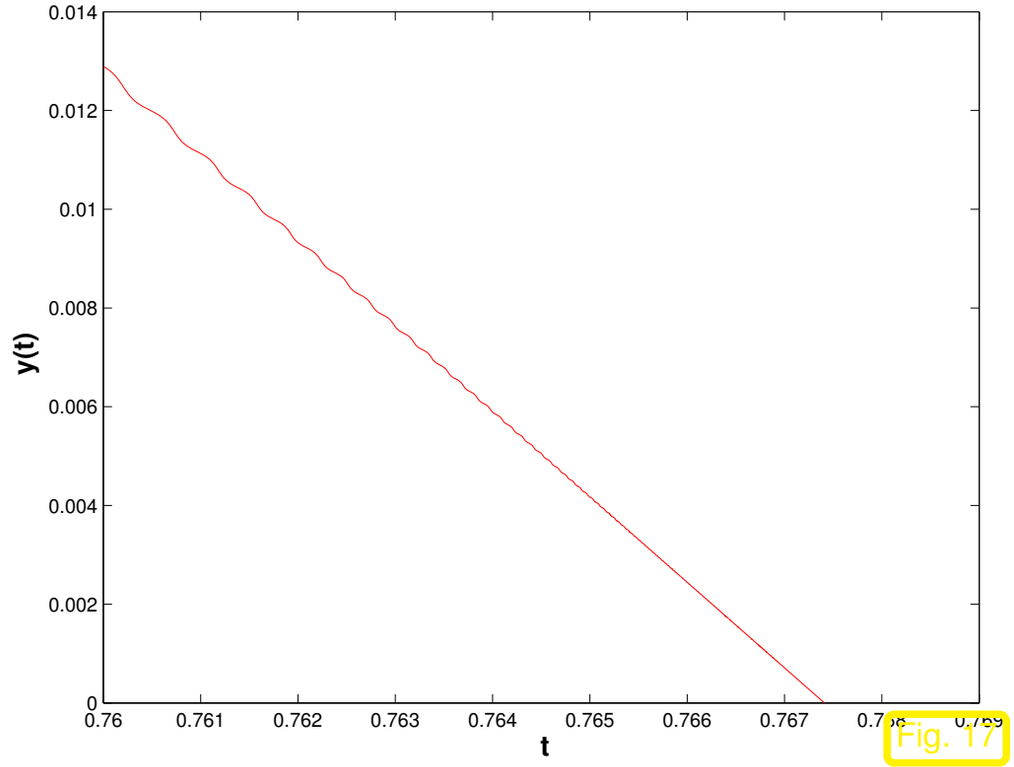


Fig. 17

◇

1.3.2 Linear IVPs [5, Sekt. 8.2]

Preparation: change of basis in state space (covariant transformation):

$$\widehat{\mathbf{y}} = \mathbf{S}^{-1}\mathbf{y} \quad , \quad \mathbf{S} \in \mathbb{R}^{d,d} \quad \text{regular matrix (independent of time).}$$

$$\mathbf{y} \text{ solves } \begin{cases} \dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y}) \,, \\ \mathbf{y}(t_0) = \mathbf{y}_0 \end{cases} \Leftrightarrow \widehat{\mathbf{y}} := \mathbf{S}^{-1}\mathbf{y} \text{ solves } \begin{cases} \dot{\widehat{\mathbf{y}}} = \widehat{\mathbf{f}}(t,\widehat{\mathbf{y}}) \,, \\ \widehat{\mathbf{y}}(t_0) = \mathbf{S}^{-1}\mathbf{y}_0 \end{cases} \text{ with } \widehat{\mathbf{f}}(t,\widehat{\mathbf{y}}) = \mathbf{S}^{-1}\mathbf{f}(t,\mathbf{S}\widehat{\mathbf{y}} \tag{1.3.12}$$

Consider a linear differential equation with constant coefficients in $\mathbb{R}^d$:

- $D = \mathbb{R}^d$, $\Omega = I \times \mathbb{R}^d$
- coefficient matrix $\mathbf{A} \in \mathbb{R}^{d,d}$
- "source term": continuous function $\mathbf{g} : I \mapsto \mathbb{R}^d$

$$\boxed{\dot{\mathbf{y}} = \mathbf{A}\mathbf{y} + \mathbf{g}(t)} \,. \tag{1.3.13}$$

Assumption: $\mathbf{A}$ diagonalizable, $\exists \mathbf{S} \in \mathbb{R}^{d,d}$ regular: $\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$, $\lambda_i \in \mathbb{C}$

- $\mathbf{g} \equiv 0$: $\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}$ (autonomous homogeneous linear differential equation, more general

P. Grohs

rev 63606, February 17, 2014

discussion in [12, Sect. 3.2.2])

$$\widehat{\mathbf{y}} := \mathbf{S}^{-1}\mathbf{y} \quad \text{solves} \quad \begin{aligned} \dot{\widehat{y}}_1 &= \lambda_1 \widehat{y}_1 \,, \\ &\vdots \\ \dot{\widehat{y}}_d &= \lambda_d \widehat{y}_d \end{aligned} \quad \Rightarrow \quad \widehat{y}_i(t) = (\mathbf{S}^{-1}\mathbf{y}_0)_i \, e^{\lambda_i t} \,, \quad t \in \mathbb{R} \,.$$

$$\blacktriangleright \quad \mathbf{y}(t) = \underbrace{\mathbf{S} \begin{pmatrix} e^{\lambda_1 t} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & e^{\lambda_d t} \end{pmatrix} \mathbf{S}^{-1}}_{\text{matrix exponential function } \exp(\mathbf{A}t)} \mathbf{y}_0 \,.$$

General definition of matrix exponential function by

"matrix exponential series": $$\exp(\mathbf{M}) = \sum_{k=0}^{\infty} \frac{1}{k!} \mathbf{M}^k \,. \tag{1.3.14}$$

Important properties: $\boxed{\text{matrix exponential function commutes with similarity transformations}}$

$$\mathbf{M} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S} \quad \Rightarrow \quad \exp(\mathbf{M}) = \mathbf{S}^{-1}\exp(\mathbf{A})\mathbf{S} \quad \forall \mathbf{A}, \mathbf{M}, \mathbf{S} \in \mathbb{C}^{d,d}, \ \mathbf{S} \text{ regular.} \tag{1.3.15}$$

- Inhomogeneous case $\dot{\mathbf{y}}(t) = \mathbf{A}\mathbf{y}(t) + \mathbf{g}(t)$ particular solution by "'variation of constants":

Ansatz: $\mathbf{y}(t) = \exp(\mathbf{A}t)\mathbf{z}(t)$ with $\mathbf{z} \in C^1(\mathbb{R}, \mathbb{R}^d)$ $\rightarrow$ [2, Thm I(5.14)]

$$\dot{\mathbf{y}}(t) = \mathbf{A}\exp(\mathbf{A}t)\mathbf{z}(t) + \exp(\mathbf{A}t)\dot{\mathbf{z}}(t) = \mathbf{A}\mathbf{y}(t) + \mathbf{g}(t) = \mathbf{A}\exp(\mathbf{A}t)\mathbf{z}(t) + \mathbf{g}(t)$$

- $\dot{\mathbf{z}}(t) = \exp(-\mathbf{A}t)\mathbf{g}(t) \quad \Rightarrow \quad \mathbf{z}(t) = \mathbf{z}(t_0) + \int_{t_0}^{t} \exp(-\mathbf{A}\tau)\mathbf{g}(\tau)\,\mathrm{d}\tau$
- $\mathbf{y}(t) = \boxed{\exp(\mathbf{A}(t-t_0))\mathbf{y}_0} + \boxed{\int_{t_0}^{t} \exp(\mathbf{A}(t-\tau))\mathbf{g}(\tau)\,\mathrm{d}\tau} =: \mathbf{\Phi}^{t_0,t}\mathbf{y}_0$.

solution of homogeneous problem

convolution with inhomogeneity

More general considerations $\rightarrow$ [2, Kap. III]:

Remark 1.3.16 (General formula for variation of constants). $\rightarrow$ [2, Thm. (11.13)]

- $\mathbf{A} : J \subset \mathbb{R} \mapsto \mathbb{R}^{d,d}$ continuous matrix function, $J \subset \mathbb{R}$ interval
- $\mathbf{g} : J \mapsto \mathbb{R}^d$ continuous
- $(s,t) \mapsto \mathbf{E}(s,t) \in \mathbb{R}^{d,d}$ describes evolution, defined by

$$\frac{\partial \mathbf{E}}{\partial t}(s,t) = \mathbf{A}(t)\mathbf{E}(s,t) \quad \forall (s,t) \in J \times J\,, \quad \mathbf{E}(s,s) = \mathbf{I}\,. \tag{1.3.17}$$

Then, the (unique $\rightarrow$ Thm. 1.3.4) solution of the non-autonomous *linear* initial value problem

$$\dot{\mathbf{y}} = \mathbf{A}(t)\mathbf{y} + \mathbf{g}(t) \quad , \quad \mathbf{y}(t_0) = \mathbf{y}_0 \ ,$$

is given by

$$\mathbf{y}(t) = \mathbf{E}(t,t_0)\mathbf{y}_0 + \int_{t_0}^{t} \mathbf{E}(t,s)\mathbf{g}(s)\,\mathrm{d}s \ , \quad t \in J \ . \tag{1.3.18}$$

Remark 1.3.19 (Significance of linear IVPs).

Linearize around a stationary point $\mathbf{f}(\mathbf{y}^*) = 0$:

$$\mathbf{y} \approx \mathbf{y}^* : \quad \mathbf{f}(\mathbf{y}) = D_{\mathbf{y}}\mathbf{f}(\mathbf{y}^*)(\mathbf{y} - \mathbf{y}^*) + O(|\mathbf{y} - \mathbf{y}^*|^2) \ ,$$

if $\mathbf{f} \in C^2$.

➢ Solutions of $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ behave *in the neighborhood of* $\mathbf{y}^*$ (qualitatively) like solutions of the linear ODE $\dot{\mathbf{y}} = D_{\mathbf{y}}\mathbf{f}(\mathbf{y}^*)\mathbf{y}$.

△

1.3.3 Sensitivity[12, Sect. 3.1]

1.3.3.1 Basic concepts

Reminder ($\rightarrow$ course "Numerical methods"):

Problem = map $\Pi : X \mapsto Y$ from data space X to sample space Y
(both endowed with metrics d_X, d_Y)

Problem is well-posed, if Π is continuous.

Condition/sensitivity of a problem:
measures how strongly perturbations in the data influence the result

(absolute) condition number: $$\kappa_{\text{abs}} := \sup_{x,x' \in X, x \neq x'} \frac{d_Y(\Pi(x), \Pi(x'))}{d_X(x, x')} . \tag{1.3.20}$$

Language use: a problem is " well conditioned", if "'$\kappa_{\text{abs}} \approx 1$"

- Whether a problem is well-posed and well conditioned depends significantly on the chosen metrics. For practical problems, these are determined by the user's needs.
- As errors in the input data are inevitable, solving problems numerically makes only sense for well-posed problems.
- The (absolute) condition number is the global Lipschitz constant of the problem map (with respect to the chosen metrics).

Asymptotic (absolute) condition number (linearized perturbation theory):

$$\kappa_{\text{abs}}^{\infty} := \lim_{\delta \to 0} \sup_{0 < d_X(x,x') < \delta} \frac{d_Y(\Pi(x), \Pi(x'))}{d_X(x, x')}$$

➡ $\kappa_{\text{abs}}^{\infty}$ measures the influence of "small perturbations"

Technique: Differential condition analysis for differentiable $\Pi : X \subset \mathbb{R}^m \mapsto Y \subset \mathbb{R}^n$,

$$\kappa_{\text{abs}}^{\infty} = \sup_{\boldsymbol{x} \in X} \|D\Pi(\boldsymbol{x})\| \ , \tag{1.3.21}$$

where $\|\cdot\| \hat{=}$ matrix norm induced by vector norm on $\mathbb{R}^m$, $\mathbb{R}^n$.

1.3.3.2 Our problem: the initial value problem

Application (of abstract concepts) on initial value problem (1.1.13):

We consider different input data, i.e., we use values or maps that have been perturbed. All other parts of the initial value problem are assumed to be unperturbed.

Scenario ❶:

- Input data $\mathbf{y}_0$ ➥ Data space $\mathbb{R}^d$, metric: Euclidean vector norm
- Result $\mathbf{y}(T)$ at final time $T > t_0$ ➥ sample space $\mathbb{R}^d$, metric: Euclidean vector norm

Scenario ❷:

- Input data $\mathbf{y}_0$ ➥ Data space $\mathbb{R}^d$, metric: Euclidean vector norm
- Result: solution function $t \in J \subset I \mapsto \mathbf{y}(t)$
  ➥ sample space $C^0(J, \mathbb{R}^d)$, metric: maximum norm $\|\cdot\|_{L^\infty(J)}$

Scenario ❸:

- Input data: Initial value $\mathbf{y}_0$ and right hand side $\mathbf{f}$
  ➥ Data space $\mathbb{R}^d \times C^1(I \times \mathbb{R}^d, \mathbb{R}^d)$, metric: Euclidean vector norm & maximum norm
- Result: Solution function $t \in J \subset I \mapsto \mathbf{y}(t)$
  ➥ sample space $C^0(J, \mathbb{R}^d)$, metric: maximum norm $\|\cdot\|_{L^\infty(J)}$

Terminology: Scenario ❶ : $\kappa_{\text{abs}}, \kappa_{\text{abs}}^\infty$ $\sim$ pointwise condition,
Scenario ❷ : $\kappa_{\text{abs}}, \kappa_{\text{abs}}^\infty$ $\sim$ interval condition

Example 1.3.22 (Condition of scalar linear initial value problems).

$$\dot{y} = \lambda y \ , \quad \lambda \in \mathbb{R} \quad , \quad y(0) = y_0 \in \mathbb{R} \ , \tag{1.3.23}$$

$$\Rightarrow \qquad y(t) = y_0 \exp(\lambda t) \ , \quad t \in \mathbb{R} \ . \tag{1.3.24}$$

(Examination for scenarios ❶ and ❷)

Pointwise condition: for final time T:

$$\kappa_{\text{abs}} = \exp(\lambda T) \qquad \begin{cases} \gg 1 & \text{for } \lambda > 0 \ , \\ \ll 1 & \text{for } \lambda < 0 \ . \end{cases} \tag{1.3.25}$$

Interval condition: in $[0,T]$:

$$\kappa_{\text{abs}} = \max\{1, \exp(\lambda T)\} \qquad \begin{cases} \gg 1 & \text{for } \lambda > 0 \ , \\ 1 & \text{for } \lambda < 0 \ . \end{cases} \tag{1.3.26}$$

$\diamond$

1.3.3.3 Well-posedness

[Beitrag von J. Mitrovic, FS 2011]

Important: Typically, input data for computations are perturbed, e.g., by rounding. Therefore, in order for the computation to make any sense, it is *essential* that the solution of the ODE depends continuously on the initial data.

Assumption: Right hand side $\mathbf{f} : I \times D \mapsto \mathbb{R}^d$ of (1.1.13) fulfills global Lipschitz condition (cmp. *local* Lipschitz condition from Def. 1.3.2)

$$\forall t \in I: \quad \exists L(t) > 0: \quad \|\mathbf{f}(t,\boldsymbol{x}) - \mathbf{f}(t,\boldsymbol{y})\| \leq L(t)\,\|\boldsymbol{x} - \boldsymbol{y}\| \quad \forall \boldsymbol{x},\boldsymbol{y} \in D \subset \mathbb{R}^d\,, \tag{1.3.27}$$

for appropriate vector norm $\|\cdot\|$ on $\mathbb{R}^d$.

Theorem 1.3.28 (Lipschitz continuous dependency on the initial value)**.**
Let $\mathbf{y}, \widetilde{\mathbf{y}}$ *be solutions of the IVP* (1.1.13) *for initial values* $\mathbf{y}_0 \in D$ *and* $\widetilde{\mathbf{y}}_0 \in D$*, respectively. Assuming* (1.3.27) *with continuous* $L(t)$*, there holds*

$$\|\mathbf{y}(t) - \widetilde{\mathbf{y}}(t)\| \leq \|\mathbf{y}_0 - \widetilde{\mathbf{y}}_0\| \cdot \exp\left(\int_{t_0}^{t} L(\tau)\,\mathrm{d}\tau\right) \quad \forall t \in I\,.$$

Auxiliary tool for the proof:

Lemma 1.3.29 (Gronwall's Lemma). $\rightarrow$ *[2, Sect. II.6], [12, Lemma 3.9]*

Let $J \subset \mathbb{R}$ be an interval, $t_0 \in J$, $u, a, \beta \in C^0(J, \mathbb{R}^+)$, and a monotonously increasing. Then, there holds

$$u(t) \leq a(|t - t_0|) + \int_{t_0}^{t} \beta(\tau) u(\tau) \, \mathrm{d}\tau \quad \Rightarrow \quad u(t) \leq a(|t - t_0|) \exp \left| \int_{t_0}^{t} \beta(\tau) \, \mathrm{d}\tau \right| .$$

[Beitrag von J. Mitrovic, FS 2011]

Proof: We only discuss the case $a = const \geq 0$.

We define: $\Psi(t) = a + \int_{t_0}^{t} \beta(\tau) u(\tau) d\tau$

$\Rightarrow \dot{\Psi}(t) = \beta(t) u(t)$.

Due to the assumption, there holds $u \leq \Psi \Rightarrow \dot{\Psi}(t) = \beta(t) u(t) \leq \beta(t) \Psi(t)$, as $u(t) \geq 0$

$\Rightarrow \frac{\dot{\Psi}(t)}{\Psi(t)} \leq \beta(t) \Rightarrow \frac{d}{dt} \log(\Psi(t)) \leq \beta(t)$

$\Rightarrow \log(\Psi(t)) \leq \int_{t_0}^{t} \beta(\tau) d\tau + C \Rightarrow u(t) \leq \Psi(t) \leq \underbrace{e^C}_{=:a} e^{\int_{t_0}^{t} \beta(\tau) d\tau}$ $\square$

➤ IVP (1.1.13) is well-posed under the assumption (1.3.27) **!**

➤ *Bound* for absolute pointwise condition number (for final time T)

$$\kappa_{\text{abs}} \leq \exp\left(\int_{t_0}^{T} L(\tau)\,\mathrm{d}\tau\right) \quad . \tag{1.3.31}$$

Remark 1.3.32 ("Gronwall-bound" for condition number).

The bound (1.3.31) is often extremely pessimistic !

Example (compare Ex. 1.3.22): for scalar linear IVP with $\lambda < 0$, pointwise condition, final time $T > 0$ (1.3.23)

$$(1.3.31) \quad \succ \quad \kappa_{\text{abs}} \leq e^{|\lambda| T} \overset{T\to\infty}{\longmapsto} \infty \qquad \longleftrightarrow \qquad \kappa_{\text{abs}} = e^{\lambda T} \overset{T\to\infty}{\longmapsto} 0 \; .$$

△

1.3.3.4 Asymptotic condition number

Scenario ❶: How do *small perturbations in the initial value* $\mathbf{y}_0$ in (1.1.13) influence the solution $\mathbf{y}(t)$?

Asymptotic absolute condition number by means of differential condition analysis, compare (1.3.21):

To do: "'Differentiate the solution of an initial value problem with respect to the initial value $\mathbf{y}_0$"
(Here, treat the time t as fixed "parameter".)

➢ Consider, with evolution $\mathbf{\Phi}^{s,t}\mathbf{y} = \mathbf{\Phi}(s,t,\mathbf{y})$

$$\left.\frac{d\mathbf{y}(t)}{d\mathbf{y}_0}\right|_{t \text{ fest}} \longleftrightarrow \frac{\partial \mathbf{\Phi}}{\partial \mathbf{y}}(t_0,t,\mathbf{y}_0) \quad \text{(partial derivative w.r.t. } \mathbf{y}\text{)} .$$

Formal procedure assuming that partial derivatives commute:

$$\frac{d}{dt}\mathbf{\Phi}^{t_0,t}\mathbf{y} = \mathbf{f}(t,\mathbf{\Phi}^{t_0,t}\mathbf{y}) \quad \text{for fixed } \mathbf{y} .$$

$$\xrightarrow{\frac{d}{d\mathbf{y}}} \quad \frac{d}{d\mathbf{y}}\left(\frac{\partial}{\partial t}\mathbf{\Phi}^{t_0,t}\mathbf{y}\right) = \frac{d}{d\mathbf{y}}\mathbf{f}(t,\mathbf{\Phi}^{t_0,t}\mathbf{y})$$

$$\frac{\partial}{\partial t}\left(\frac{\partial \mathbf{\Phi}}{\partial \mathbf{y}}(t_0,t,\mathbf{y})\right) = \frac{d}{d\mathbf{y}}\mathbf{f}(t,\mathbf{\Phi}^{t_0,t}\mathbf{y}) = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t,\mathbf{\Phi}^{t_0,t}\mathbf{y})\frac{d}{d\mathbf{y}}\mathbf{\Phi}^{t_0,t}\mathbf{y} .$$

(Assumption: $\mathbf{f}$ continuously differentiable w.r.t. $\mathbf{y}$)

The propagation matrix (Wronski-matrix),

$$\mathbf{W}(t;t_0,\mathbf{z}) := \frac{\partial}{\partial \mathbf{y}} \mathbf{\Phi}^{t_0,t}\mathbf{y}_{|\mathbf{y}=\mathbf{z}} \in \mathbb{R}^{d,d} \,, \tag{1.3.33}$$

for the IVP (1.1.13) fulfills the IVP

$$\text{variational equation} \quad \begin{aligned} &\frac{d}{dt}\mathbf{W}(t;t_0,\mathbf{y}_0) = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t,\mathbf{\Phi}^{t_0,t}\mathbf{y}_0)\mathbf{W}(t;t_0,\mathbf{y}_0) \,, \\ &\mathbf{W}(t_0;t_0,\mathbf{y}_0) = \mathbf{I} \,. \end{aligned} \tag{1.3.34}$$

P. Grohs

rev 63606, February 17, 2014

Note: variational equation = linear differential equation on state space $D = \mathbb{R}^{d,d}$

☞ Matrix differential equation of the form

$$\dot{\mathbf{W}} = \mathbf{A}(t)\mathbf{W} \quad \text{with} \quad \mathbf{A}(t) = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t,\mathbf{y}(t)) \,,$$

where $\frac{\partial \mathbf{f}}{\partial \mathbf{y}} \hat{=}$ Jacobian matrix, depending on $(t,\mathbf{y})$,
$\mathbf{y}(t) \hat{=}$ solution of IVP $\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y})$, $\mathbf{y}(t_0) = \mathbf{y}_0$

➢ To solve the variational equation, one also needs to solve the corresponding initial value problem!

$$\mathbf{y}_0 \leftarrow \mathbf{y}_0 + \delta\mathbf{y}_0 \quad \succ \quad \delta\mathbf{y}(t) \approx \mathbf{W}(t;t_0,\mathbf{y}_0)\delta\mathbf{y}_0 \quad \text{for “small } \delta\mathbf{y}_0\text{”}$$

► Interval asymptotic condition number of the IVP (1.1.13) on $[t_0, T]$ (w.r.t norm $\|\cdot\|$ on $\mathbb{R}^d$):

$$\kappa_{\text{abs}}^{\infty} := \max\{\|\mathbf{W}(t;t_0,\mathbf{y}_0)\| : t_0 \leq t \leq T\} \ . \tag{1.3.35}$$

[Beitrag von J. Mitrovic, FS 2011]

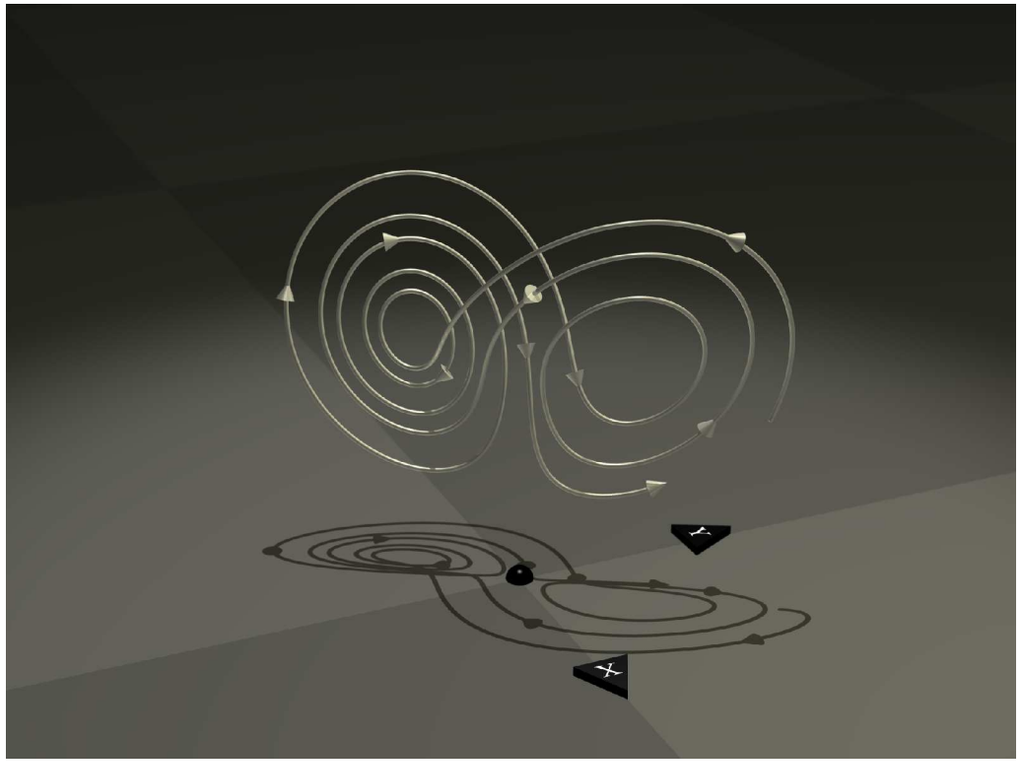
Remark: The propagation matrix (Wronski matrix) $\mathbf{W}$ measures the relative changes in the solution caused by small perturbations in the initial values.

1.3.3.5 Ill-conditioned IVPs

Example 1.3.36 (Lorenz system). $\rightarrow$ [32, 26]

Autonomous differential equation, $D = \mathbb{R}^3$, $\sigma, \rho, \beta \in \mathbb{R}^+$:

$$\begin{aligned} \dot{x} &= \sigma(y - x) \;, \\ \dot{y} &= x(\rho - z) - y \;, \\ \dot{z} &= xy - \beta z \;. \end{aligned} \tag{1.3.37}$$



P. Grohs

rev 63606, February 17, 2014

Listing 1.3: Numerical integration of Lorenz system

```
function lorenzplot(rho,sigma,beta)

% MATLAB script for plotting 3D trajectories of the Lorenz system for
  Ex. 1.3.36
% Arguments:  parameters of the Lorenz system (1.3.37):  ρ, σ, β.

% Default paramters
if (nargin < 3), rho=28; sigma = 10; beta = 8/3; end

```

```
% function handle for right hand side of the Lorez system (1.3.37)
f = @(t,y) ([sigma*(y(2)-y(1));rho*y(1) - y(2) -
  y(1)*y(3);y(1)*y(2) - beta*y(3)]);

y0 = [8 9 9.5]; ystart = y0; % initial conditions
ts = [0 20]; % Time for simulation

% Numerical integration of Lorenz system using MATLAB standard integrator,
% see Rem. 1.2.4
opts = odeset('reltol',1E-10,'abstol',1E-10,'stats','on');
[t,y] = ode45(f,ts,y0,opts);
y0(3) = y0(3) + 1.0E-5; % Slight perturbation of initial value
[tt,yt] = ode45(f,ts,y0,opts);

% 3D plot of trajectories
figure('name','Lorenz'); hold on;
plot3(y(:,1),y(:,2),y(:,3),'r-');
plot3(yt(:,1),yt(:,2),yt(:,3),'b-');
xlabel('{\bf x}','fontsize',14);
ylabel('{\bf y}','fontsize',14);
zlabel('{\bf z}','fontsize',14);
title(sprintf('\\sigma = %d, \\rho = %d, \\beta =
  %d',sigma,rho,beta));
```

```
view(45,15); grid on;
print -depsc2 'lorenz.eps';
```

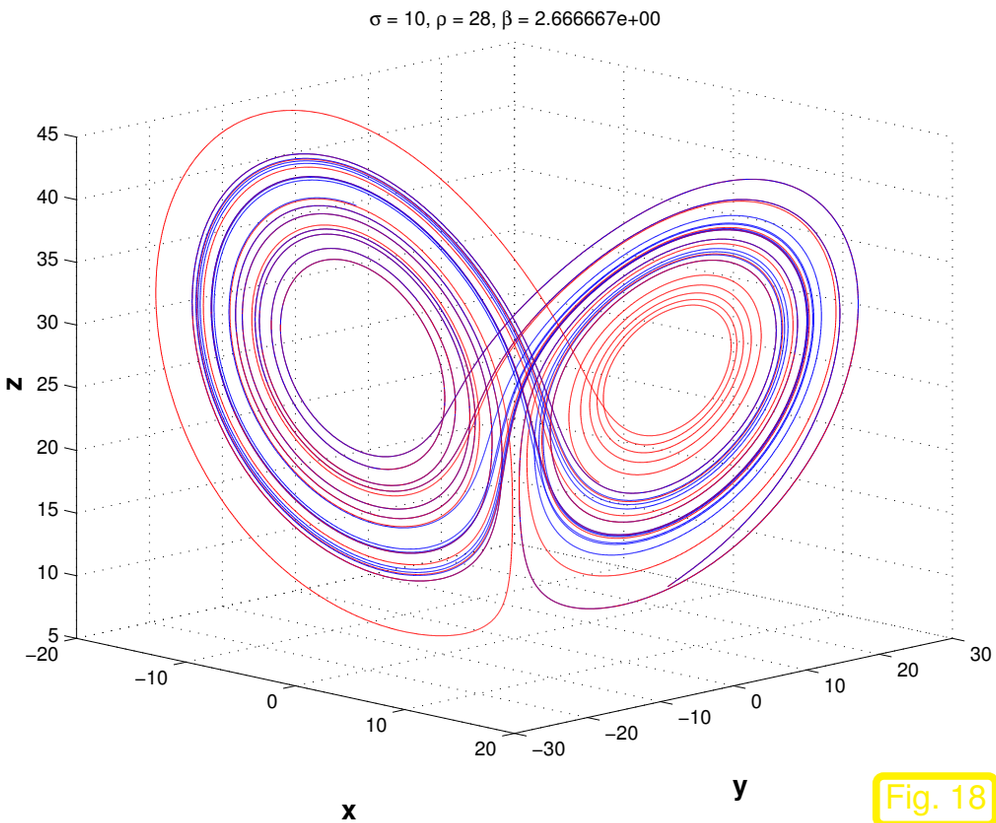


Fig. 18

— : initial value $\mathbf{y}_0 := (8, 9, 9.5)^T$

— : initial value $\mathbf{y}_0 := (8, 9, 9.5+10^{-5})^T$

From the theory of

real dynamic systems [26]:

Lorenz system is a chaotic system

Exponential divergence of bounded trajectories (Bound (1.3.31) holds for small T !)

Tiny perturbations in initial values

➤ completely different states after "exponentially short time". ◇

Goal of numerical simulations of chaotic dynamic systems:

Identification of the "typical" behavior of trajectories

Essential: Correct treatment of conservation quantities, e.g., total energy ($\rightarrow$ Invariant, Def. 1.2.7)

Example 1.3.38 (Double pendulum).

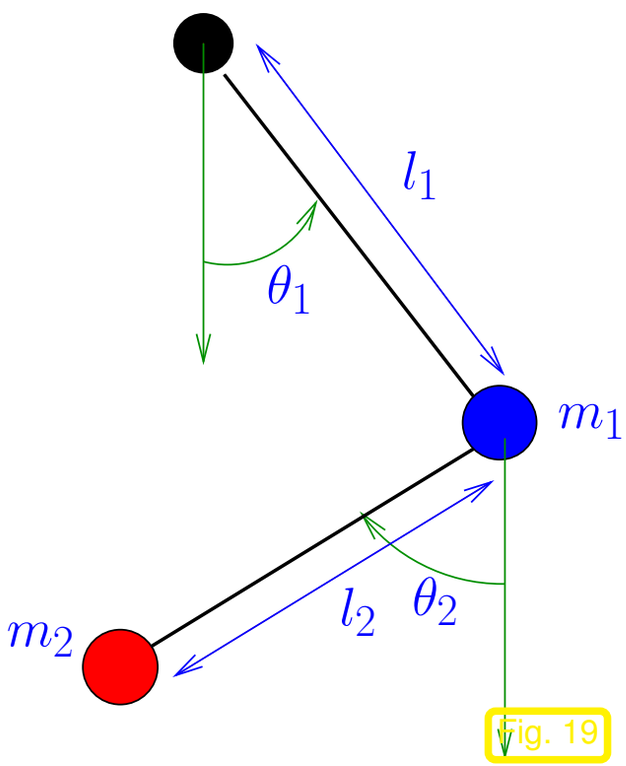


Fig. 19

$\triangleleft$ (Mathematical) double pendulum with fixed attachment and massless limbs.

Minimal coordinates: angular displacement θ_1, θ_2

configuration space $D = [0, 2\pi[^2$ (torus)

Hamiltonian ($\rightarrow$ Def. 1.2.20) = sum of kinetic and potential energy:

$$H(p_1, p_2, \theta_1, \theta_2) = \frac{l_2^2 m_2 p_1^2 + l_1^2 (m_1 + m_2) p_2^2 - 2 m_2 l_1 l_2 p_1 p_2 \cos(\theta_1 - \theta_2)}{2 l_1^2 l_2^2 m_2 (m_1 + \sin^2(\theta_1 - \theta_2) m_2)} - m_2 g l_2 \cos\theta_2 - (m_1 + m_2) g l_1 \cos\theta_1 \ .$$

Observation (in experiment and simulation):

Movement of pendulum depends extremely sensitively on the initial conditions

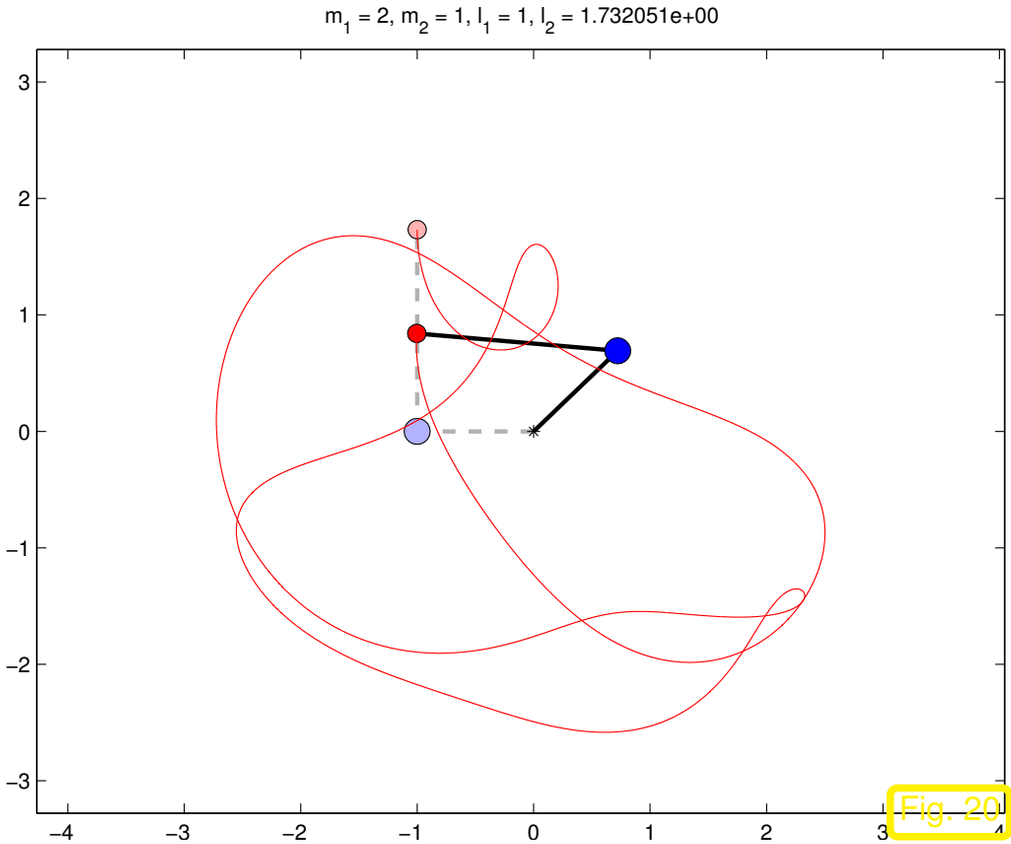


Fig. 20

$\theta_1(0) = -\frac{\pi}{2}$, $\theta_2(0) = \pi$, $p_1(0) = p_2(0) = 0$

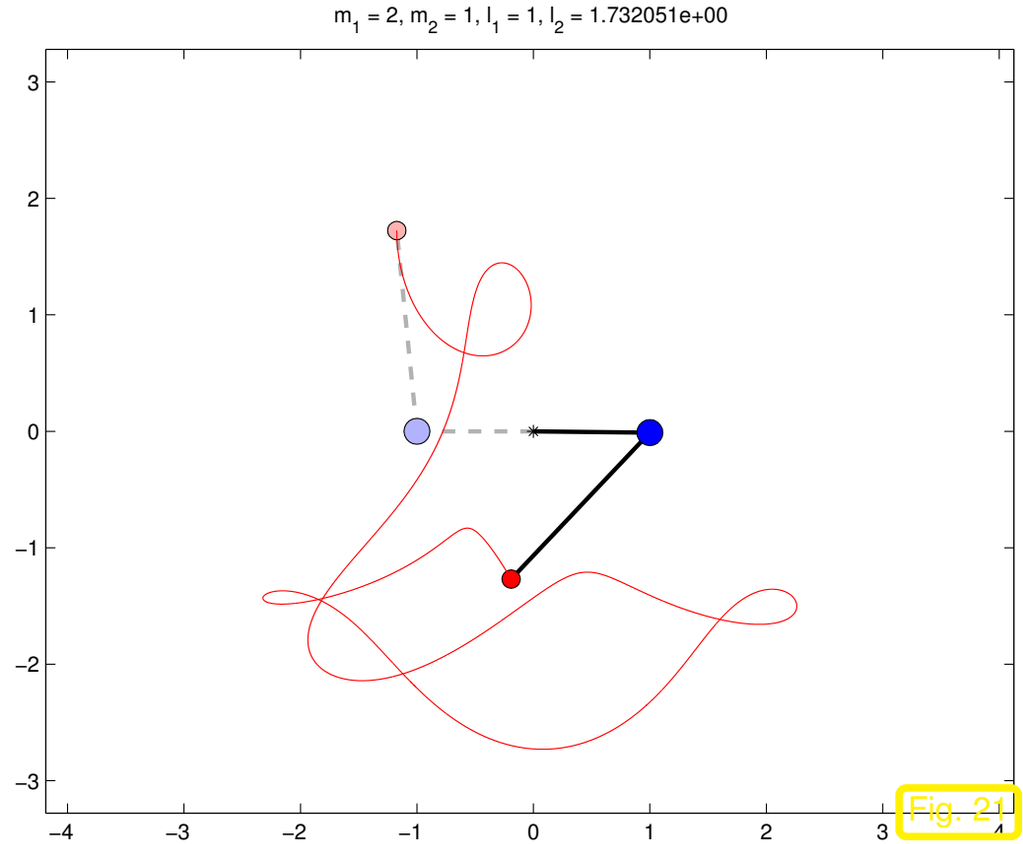


Fig. 21

$\theta_1(0) = -\frac{\pi}{2}$, $\theta_2(0) = \pi + 10^{-2}$,
$p_1(0) = p_2(0) = 0$

[Simulation, MATLAB, `ode45`, time $[0, 20]$, step size 10^{-3}]

$\diamond$

P. Grohs

rev 63606, February 17, 2014

1.4 Polygonal line methods

Given:
- Right hand side $\mathbf{f} : \Omega \mapsto \mathbb{R}^d$, locally Lipschitz continuous ($\rightarrow$ Def. 1.3.2) on augmented state space $\Omega := I \times D \subset \mathbb{R}^{d+1}$
- Initial values $\mathbf{y}_0 \in D$ at initial time t_0

Thm. 1.3.4 (Peano & Picard-Lindelöf) ➤ existence & uniqueness of solutions ($\rightarrow$ Def. 1.1.14) of the IVP

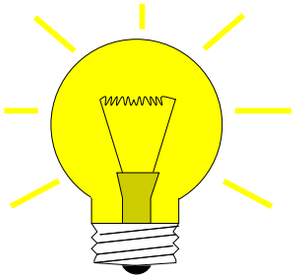
$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}) \quad , \quad \mathbf{y}(t_0) = \mathbf{y}_0 \; . \tag{1.4.1}$$

Goal:
- ☞ approximation of $\mathbf{y}(T)$ at final time $T \in J(t_0, y_0)$ $\triangleright$ $\mathbf{y}_h(T)$.
- ☞ approximation of the function $t \mapsto \mathbf{y}(t)$, $t \in [t_0, T]$, $T \in J(t_0, \mathbf{y}_0)$ $\triangleright$ $t \mapsto \mathbf{y}_h(t)$.

P. Grohs

rev 63606, February 17, 2014

1.4.1 The explicit Euler method (Euler 1768)



Idea: ❶ "Feel your way" in time: (1.4.1) = composition of IVPs for *given* small time intervals $[t_{k-1}, t_k]$, $k = 1, \dots, N$, $t_N := T$

❷ Approximate the time-local solution curves by tangents in the current initial point.

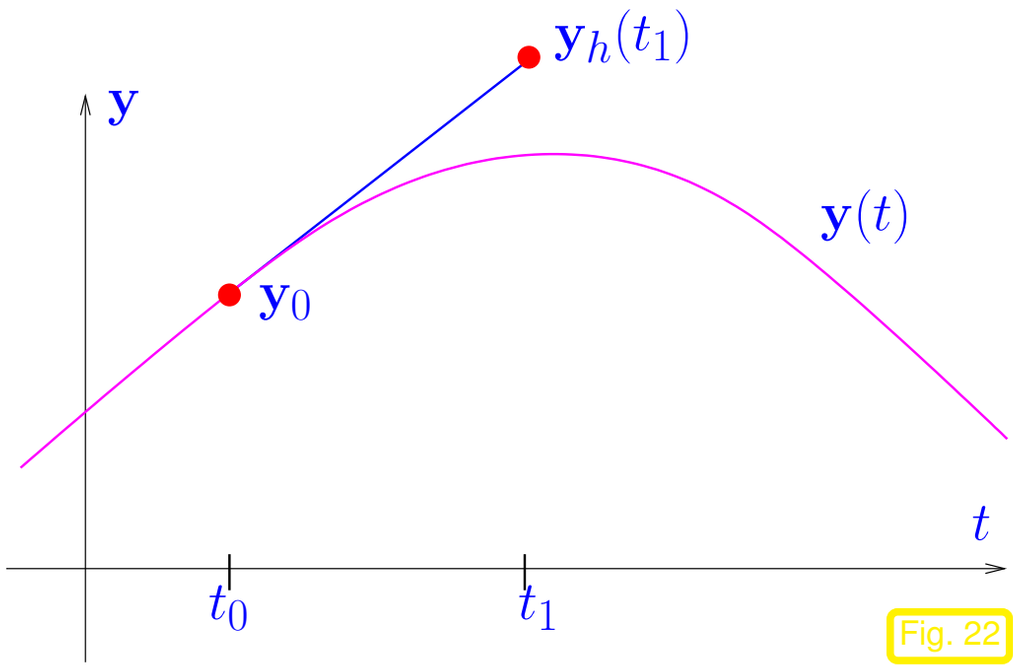


Fig. 22

Explicit Euler method

"Euler's polygonal line method"

◁ First step of explicit Euler method ($d = 1$):

Slope of tangent $= f(t_0, \mathbf{y}_0)$

$\mathbf{y}_h(t_1)$ is the initial value for the next step !

In formulae: approximations for $\mathbf{y}(t_k)$ generated by the explicit Euler method satisfy the recursion

$$\boxed{\mathbf{y}_{k+1} := \mathbf{y}_h(t_{k+1}) = \mathbf{y}_h(t_k) + h_k \mathbf{f}(t_k, \mathbf{y}_h(t_k)) \ , \quad k = 0, \dots, N-1 \quad ,} \tag{1.4.2}$$

with local (time) step size $h_k := t_{k+1} - t_k$.

✎ Alternative notation: $\mathbf{y}_k := \mathbf{y}_h(t_k)$

Illustration: explicit Euler method

initial value problem for Riccati-differential equation, compare Ex. 1.1.3

$$\dot{y} = y^2 + t^2 \ . \qquad (1.1.4)$$

$y_0 = \frac{1}{2}$, $t_0 = 0$, $T = 1$, ▷

constant time steps $h = 0.2$

$\mapsto \hat{=}$ slope field of the Riccati-differential eq.

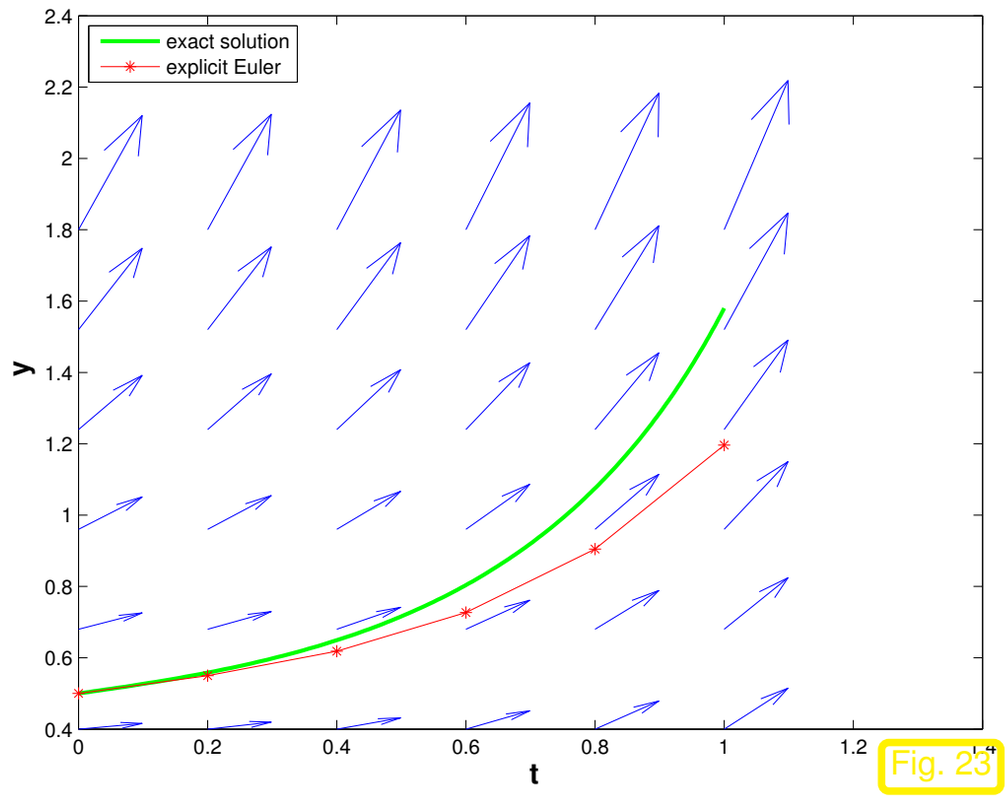


Fig. 23

P. Grohs

rev 63606, February 17, 2014

Remark 1.4.3 (Explicit Euler method as difference method).

(1.4.2) by approximating the derivative $\frac{d}{dt}$ by forward difference quotients on time grid $\mathcal{G} := \{t_0, t_1, \dots, t_N\}$:

$$\dot{\mathbf{y}} = f(t, \mathbf{y}) \quad \longleftrightarrow \quad \frac{\mathbf{y}_h(t_{k+1}) - \mathbf{y}_h(t_k)}{h_k} = f(t_k, \mathbf{y}_h(t_k)) \,, \quad k = 0, \dots, N-1 \,.$$

△

Question: How accurate is the approximate solution **?**

Example 1.4.4 ((Velocity of) Convergence of the explicit Euler method).

Listing 1.4: Generation of error curves for explicit Euler method

```
function err = eulerConvergence(odefun,tspan,y0v,N)
% MATLAB function computing the error (at final time) of the explicit Euler
  method (1.4.2)
% Arguments:
% odefun = @(t,y):  handle to function returning a vector
```

```
% tspan = [t0 T]: initial and final time
% y0v ≙ array of initial values
% N ≙ vector containing numbers of steps.  For each the error is returned

err = []; l = 1; % Initialize error array

for y0 = y0v
% Compute 'exact' solution
  [t,y] =
    ode45(odefun,tspan,y0,odeset('reltol',1E-11,'abstol',1E-11));

% Compute Euler solutions
  erri = [];
  for n=N
    h = (tspan(2)-tspan(1))/n; % uniform timestep size
    t_eul = tspan(1);          % initial time
    y_eul = y0;                % intialize iteration
    for k=1:n
      y_eul = y_eul + h*odefun(t_eul,y_eul); % see (1.4.2)
      t_eul = t_eul + h;                     % increment time
    end
    erri = [erri,norm(y(end,:)-y_eul)]; % record error
  end
```

```
  err = [err;erri]; % assemble matrix of error values
  leg{l} = sprintf('y0 = %3.2f',y0);
  l = l+1;
end

% Plot error curves in log-log scale to discern alegebraic convergence →
  Def. 1.4.5
figure('name','erreul');
ts = (tspan(2)-tspan(1))./N;
loglog(ts,err,'-+'); hold on;
loglog(ts,10*ts,'k-');
xlabel('{\bf timestep h}','fontsize',14);
ylabel('{\bf error (Euclidean norm)}','fontsize',14);
leg{l} = 'O(h)'; legend(leg,'location','southeast');
```

- IVP for Riccati differential equation (1.1.4) on $[0,1]$
- Explicit Euler method (1.4.2) with uniform time step $h = 1/n$, $n \in \{5, 10, 20, 40, 80, 160, 320, 640\}$.
- Error $\quad \mathrm{err}_h := |y(1) - y_h(1)|$

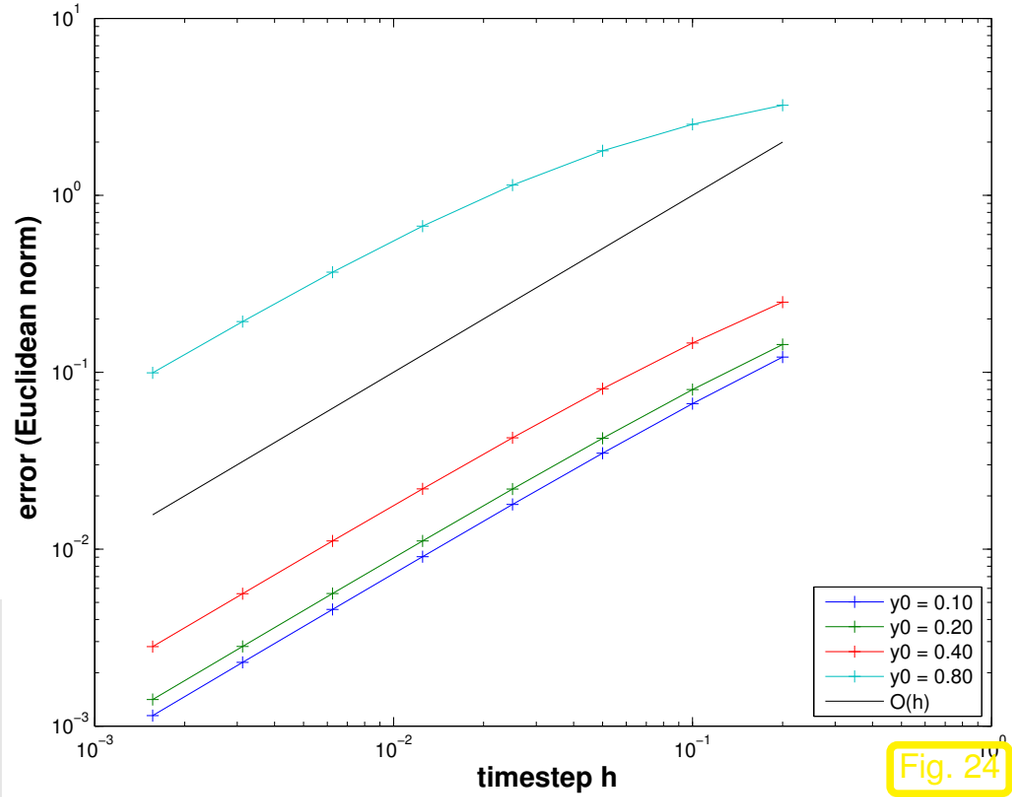


Fig. 24

Observation:

Algebraic convergence $\quad \mathrm{err}_h = O(h)$

◇

✎ Notation "Landau-O":

$$e(h) = O(g(h)) \quad \text{for } h \to 0 \quad :\Leftrightarrow \quad \exists h_0 > 0,\ C > 0\colon \quad |e(h)| \le C g(h) \quad \forall 0 \le h \le h_0 \ .$$

Definition 1.4.5 (Types of convergence)**.**

Let err_h *be the discretization error of a method for discretization parameter/step size* h, $h > 0$.

$\mathrm{err}_h = O(h^\alpha)$ $:\Leftrightarrow$ *Algebraic convergence of the order* $\alpha > 0$

$\mathrm{err}_h = O(\exp(-\beta h^{-\gamma}))$, $:\Leftrightarrow$ *exponential convergence, if* $\beta, \gamma > 0$

Error plots for algebraic convergence ($h_i = (3/2)^{-i}$, $i = 1, \ldots, 10$)

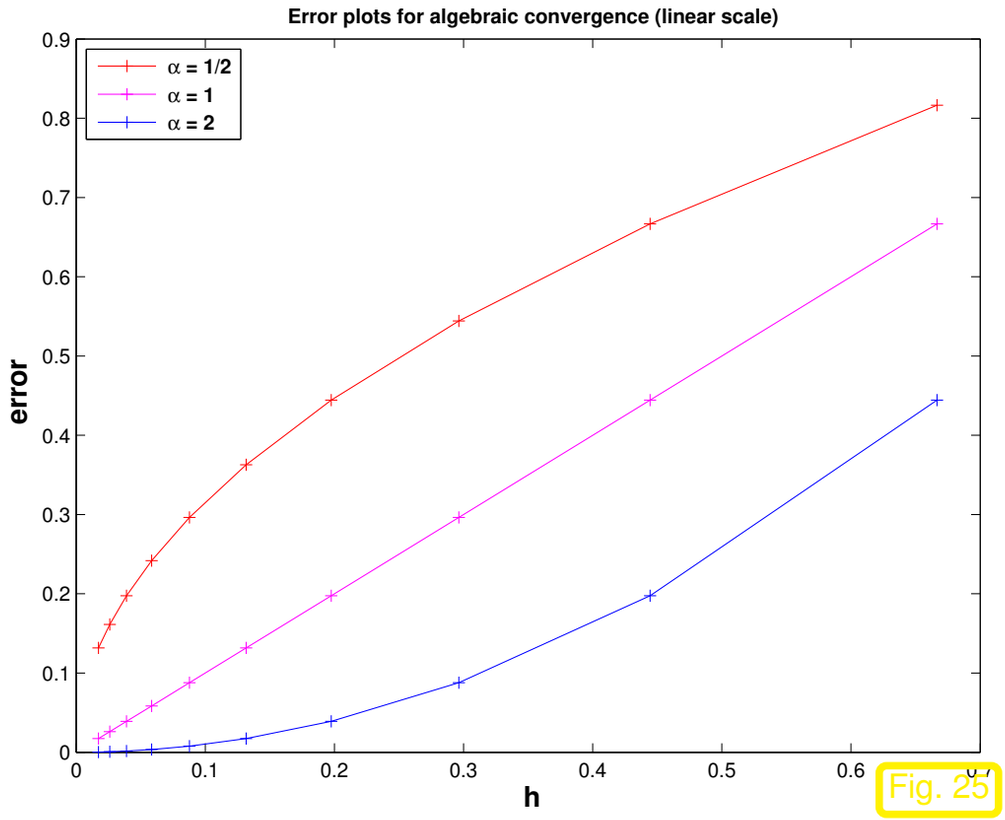


Fig. 25

linear scales

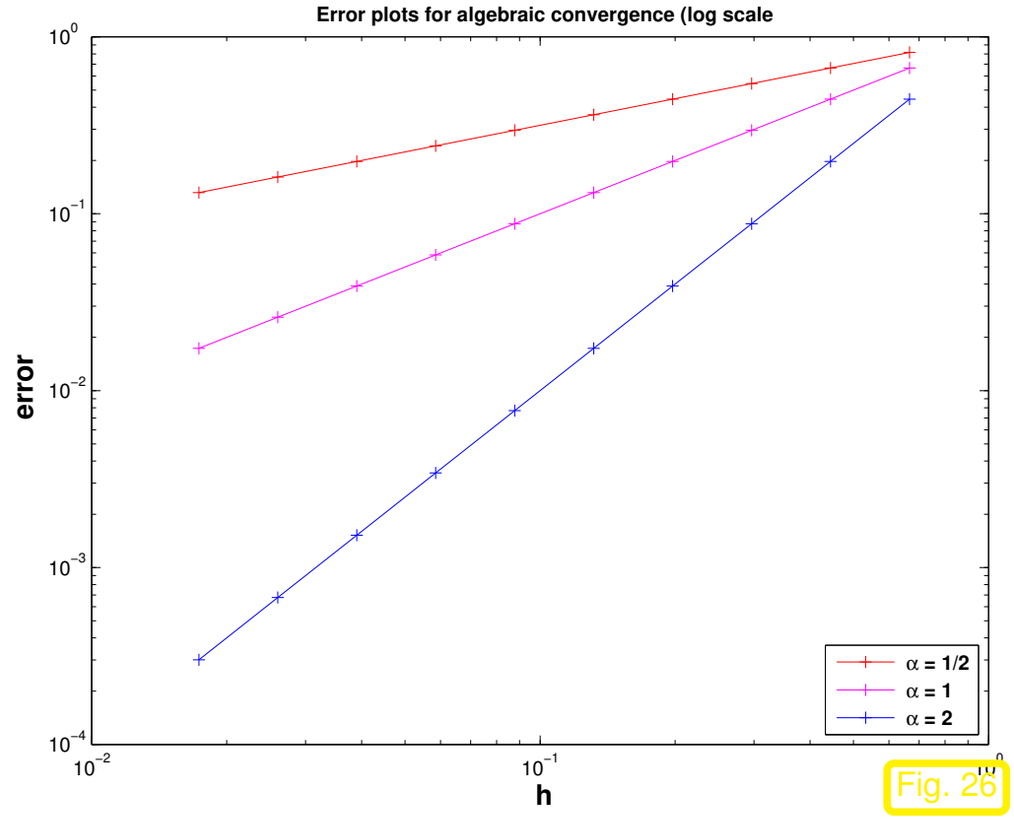


Fig. 26

logarithmic scales

Error plots for exponential convergence $(h_i = (3/2)^{-i},\ i = 1, \ldots, 10)$

P. Grohs

rev 63606, February 17, 2014

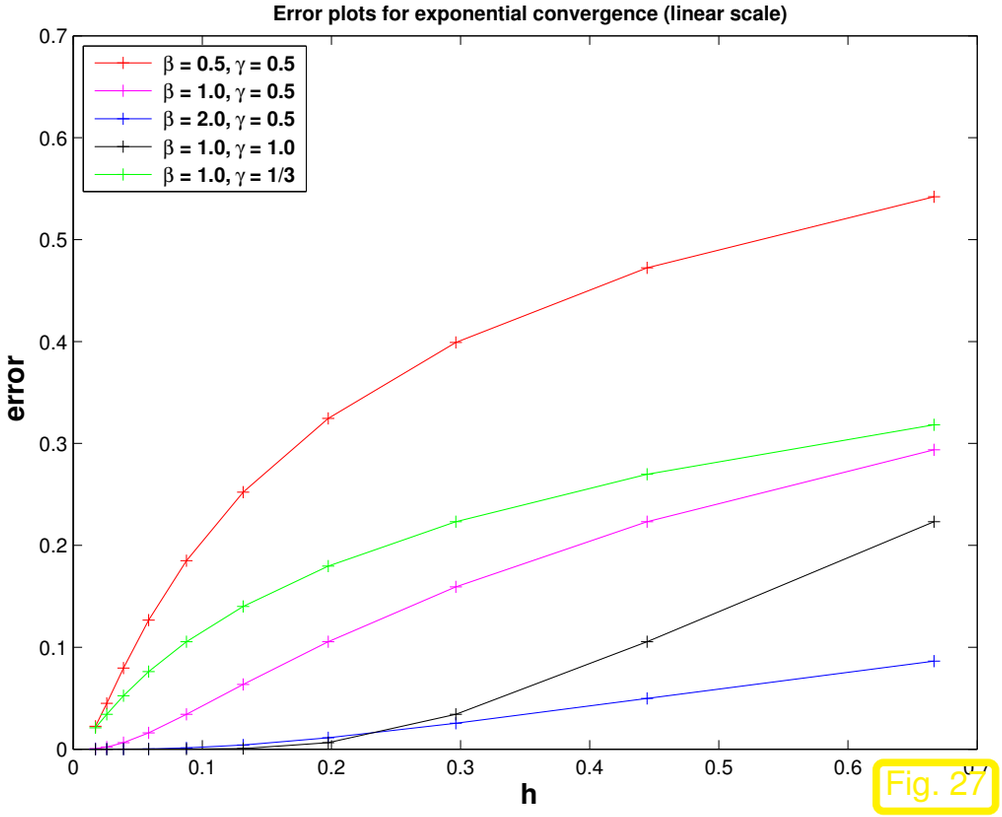


Fig. 27

linear scales

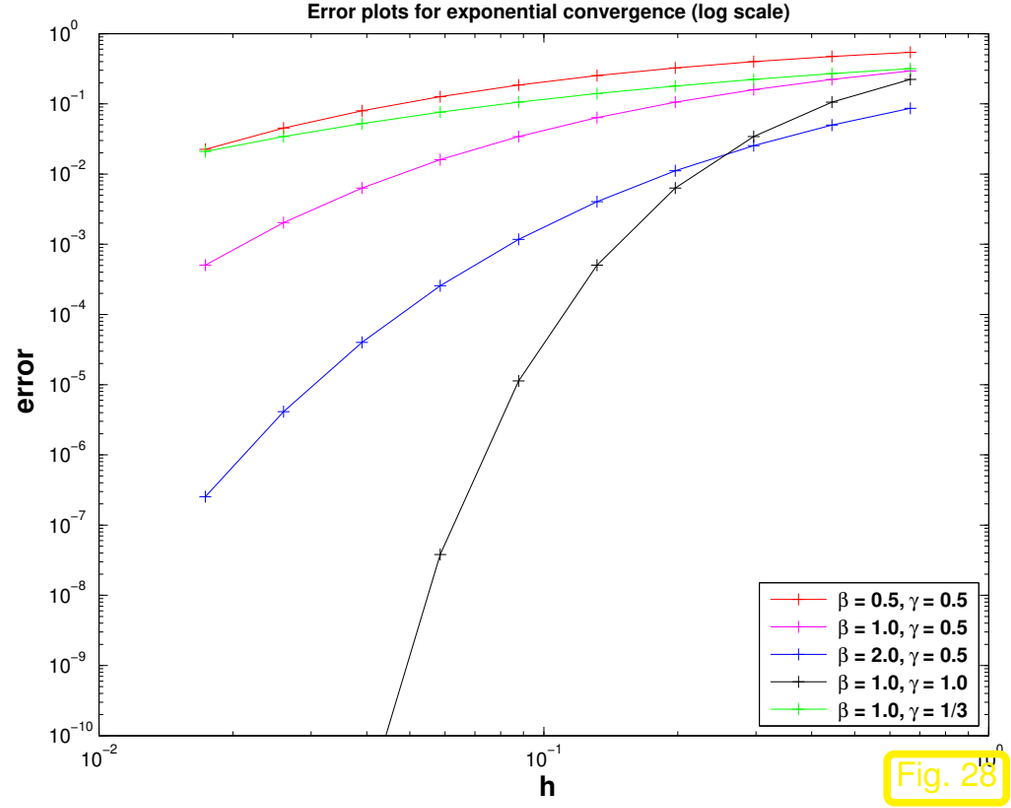
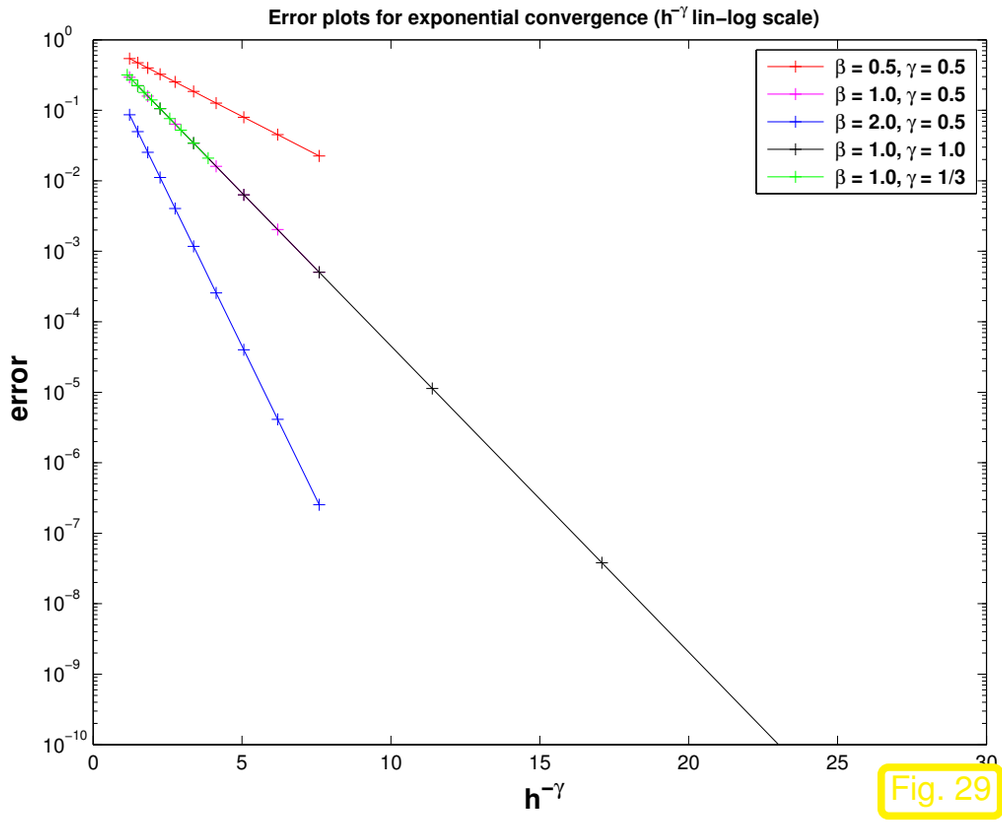


Fig. 28

logarithmic scales

P. Grohs

rev 63606, February 17, 2014



◁ $(h_i^{-\gamma}, \epsilon_i)$ ($h_i \hat{=}$ step sizes, $\epsilon_i \hat{=}$ corresponding discretization error) are on the line with slope $-\beta$

Example 1.4.9 (Explicit Euler method for logistic differential equations).

- initial value problem for logistic differential equations, compare Ex. 1.2.1

$$\dot{y} = \lambda y(1-y) \quad , \quad y(0) = 0.01 \ .$$

- Explicit Euler method (1.4.2) with uniform time step $h = 1/n$, $n \in \{5, 10, 20, 40, 80, 160, 320, 640\}$.
- Error at final time $T = 1$

P. Grohs

rev 63606, February 17, 2014

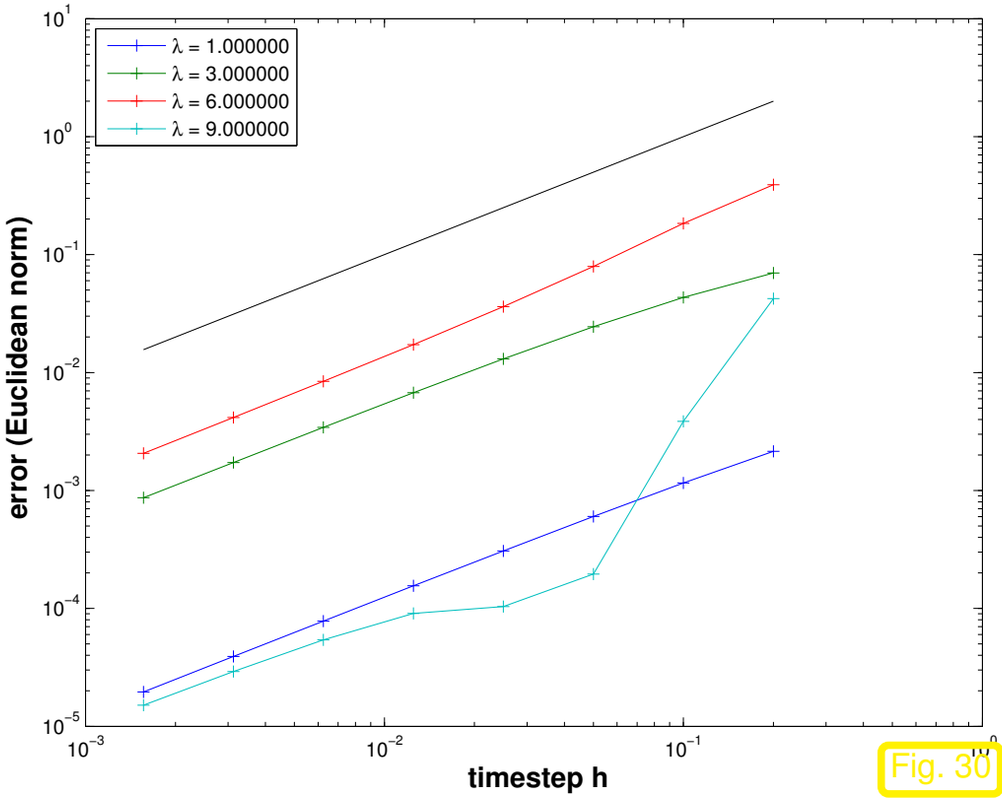


Fig. 30

λ small: $O(h)$-convergence (asymptotic)

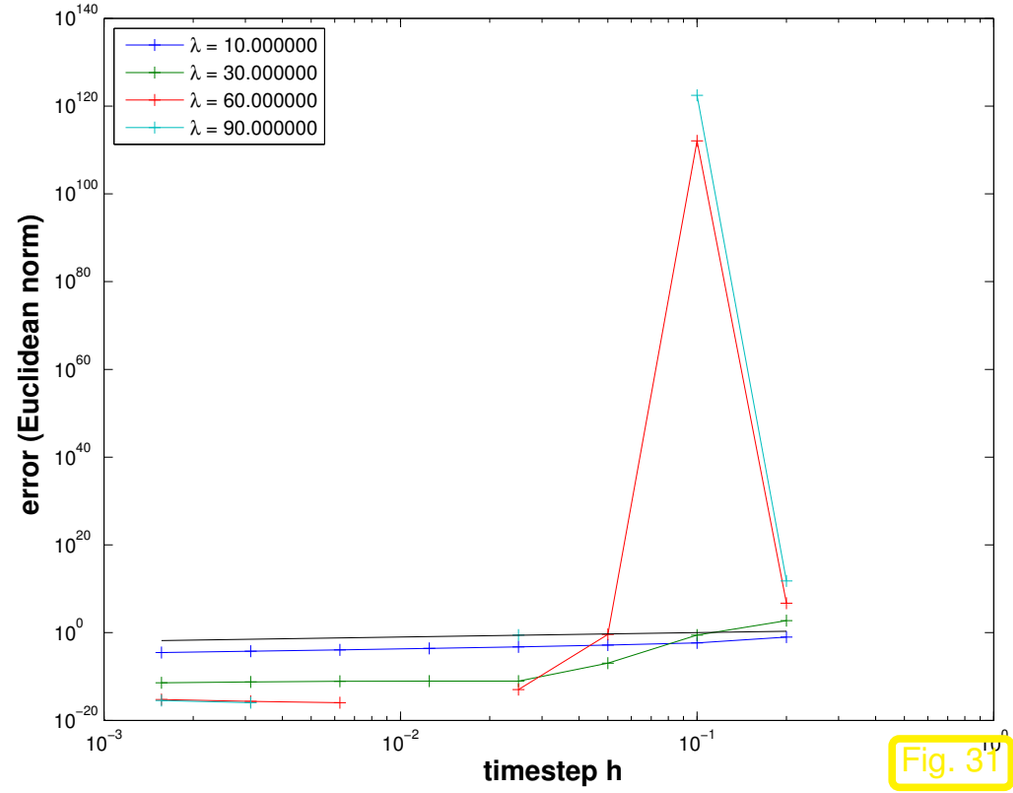


Fig. 31

λ large: explosion of y_k for large time step sizes h

P. Grohs

rev 63606, February 17, 2014

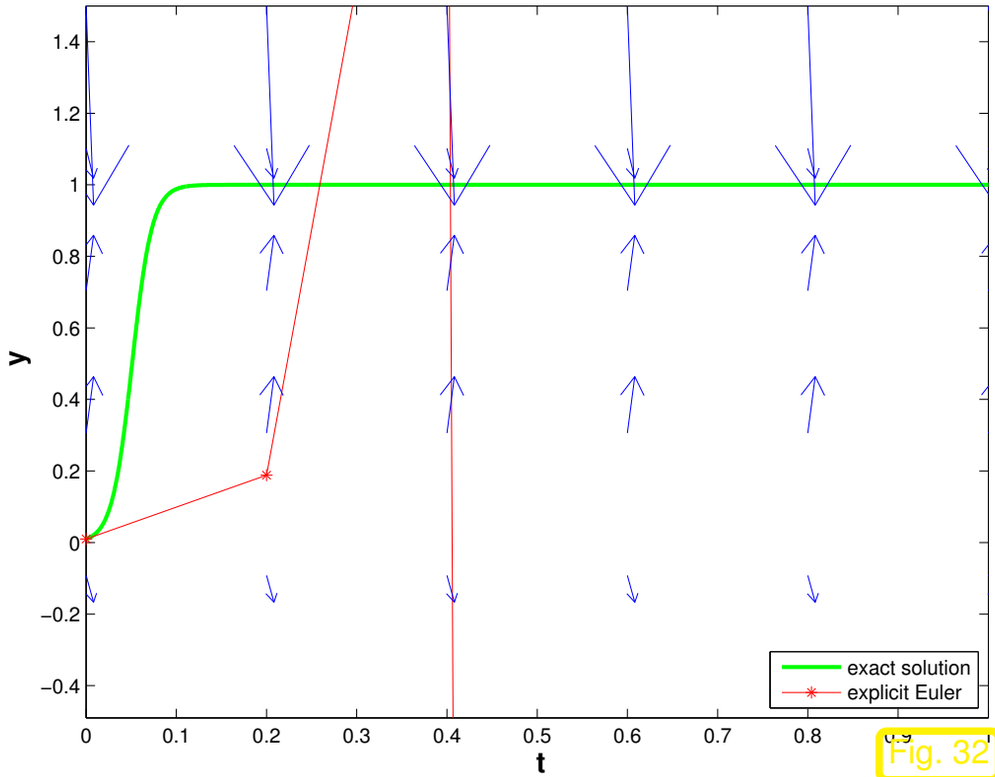


Fig. 32

◁ $\lambda = 90$, — $\hat{=}$ exact solution, — $\hat{=}$ explicit Euler

y_k overshoot the strongly attractive fixed-point $y = 1$.

➥ Observation: exponentially growing oscillation of y_k

◇

P. Grohs

rev 63606, February 17, 2014

Insight through analysis of model problem: most simple differential equation with strongly attractive fixed-point $y = 0$

Homogeneous linear scalar differential eq., Sect. 1.3.2 : $\dot{y} = f(y) := \lambda y$, $\lambda < 0$. (1.4.10)

$$\dot{y} = \lambda y \quad , \quad y(0) = y_0 \quad \Rightarrow \quad y(t) = y_0 \exp(\lambda t) \to 0 \quad \text{for} \quad t \to \infty \ . \qquad (1.4.11)$$

Recursion of explicit Euler method for (1.4.10) (uniform time step size $h > 0$)

$$\text{(1.4.2) for } f(y) = \lambda y: \qquad y_{k+1} = y_k(1 + \lambda h) \ . \tag{1.4.12}$$

$$\blacktriangleright \qquad y_k = y_0(1+\lambda h)^k \quad \Rightarrow \quad |y_k| \to \begin{cases} 0 & \text{, if } \lambda h > -2 \quad \text{(qualitatively correct)} \ , \\ \infty & \text{, if } \lambda h < -2 \quad \text{(qualitatively wrong)} \ . \end{cases}$$

More extensive treatment of the analysis of model problems in Sect. 3.1.

1.4.2 The implicit Euler method

How can we avoid that the explicit Euler method overshoots for strongly attractive fixed-points and large time steps**?**

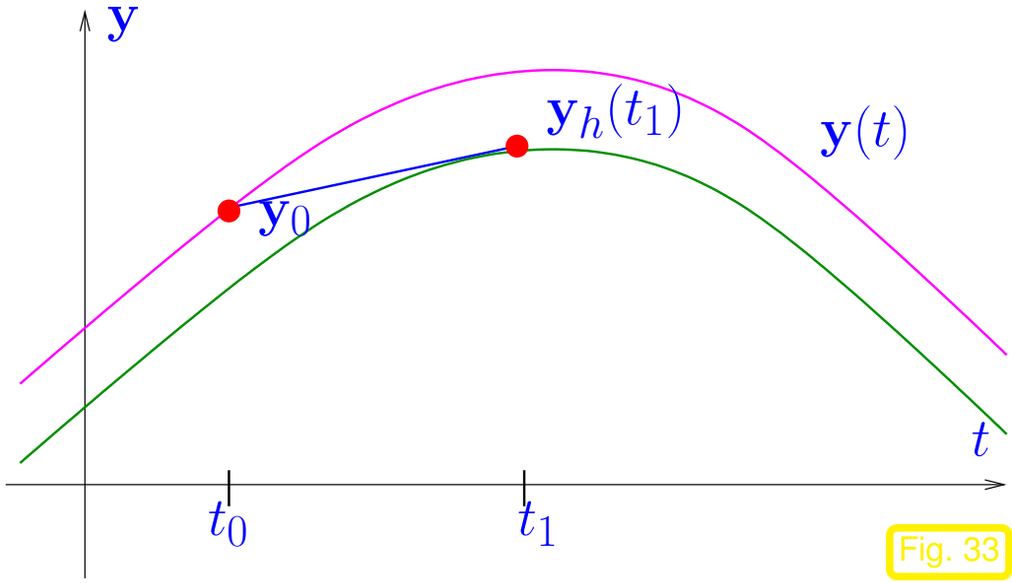


Fig. 33

Idea: approximate solution at (t_0, y_0) on $[t_0, t_1]$ by

- line through (t_0, y_0)
- with slope $f(t_1, y_1)$

◁ — $\hat{=}$ solution curve through (t_0, y_0),
— $\hat{=}$ solution curve through (t_1, y_1),
— $\hat{=}$ tangent to — at (t_1, y_1).

Apply to small time intervals $[t_0, t_1], [t_1, t_2], \ldots, [t_{N-1}, t_N]$ ➤ implicit Euler method

▶ approximations for $\mathbf{y}(t_k)$ generated by the implicit Euler method satisfy

$$\boxed{\mathbf{y}_{k+1} := \mathbf{y}_h(t_{k+1}) = \mathbf{y}_h(t_k) + h_k \mathbf{f}(t_{k+1}, \mathbf{y}_{k+1}) \;, \quad k = 0, \ldots, N-1} \;, \tag{1.4.13}$$

with local (time) step size $h_k := t_{k+1} - t_k$.

Note: (1.4.13) requires solution of a (potentially non-linear) equation for $\mathbf{y}_{k+1}$!
(➤ terminology “implicit”)

[Beitrag von J. Mitrovic, FS 2011]

Remark: The implicit Euler method

- has order of convergence 1, i.e., the error behaves like $O(h)$, compare Section 2.1.3.
- is more suitable for *stiff* initial value problems than the explicit Euler method, compare section 3.1.

Remark 1.4.14 (Implicit Euler method as difference method).

(1.4.13) by approximation of the time derivative $\frac{d}{dt}$ by backward difference quotients on time grid $\mathcal{G} := \{t_0, t_1, \dots, t_N\}$:

$$\dot{\mathbf{y}} = f(t, \mathbf{y}) \quad \longleftrightarrow \quad \frac{\mathbf{y}_h(t_{k+1}) - \mathbf{y}_h(t_k)}{h_k} = f(t_{k+1}, \mathbf{y}_h(t_{k+1})) \ , \quad k = 0, \dots, N-1 \ .$$

△

Example 1.4.15 (Implicit Euler method for logistic differential equations). → Ex. 1.4.9

Repeat numerical experiments from example 1.4.9 for implicit Euler method (1.4.13):

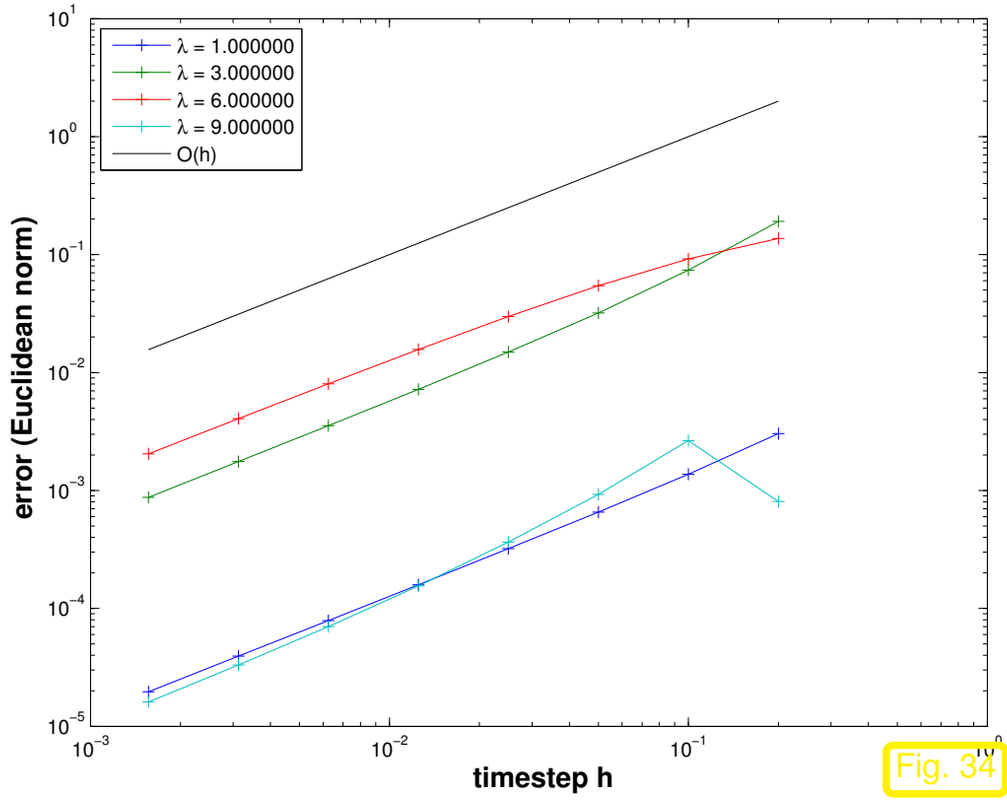


Fig. 34

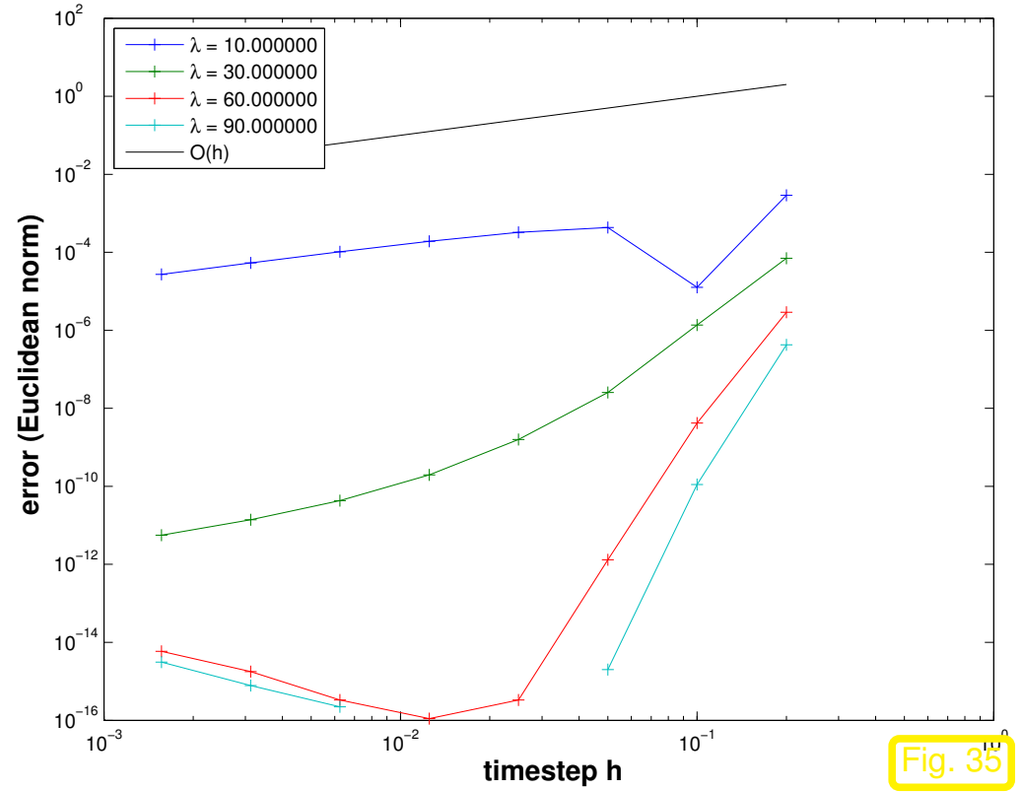


Fig. 35

λ small: $O(h)$ convergence (asymptotic)

λ large: stable for all step sizes h **!** ◇

P. Grohs

rev 63606, February 17, 2014

Analysis of model problem (like in Section 1.4.1):

(1.4.13) for $f(y) = \lambda y$: $$y_{k+1} = y_k \frac{1}{1-\lambda h} \,. \tag{1.4.16}$$

$$y_k = y_0 \left(\frac{1}{1-\lambda h}\right)^k \quad \Rightarrow \quad |y_k| \to \begin{cases} 0 & \text{, if } \lambda h < 0 \quad \text{(qualitatively correct)} \ , \\ \infty & \text{, if } 0 < \lambda h < 1 \quad \text{(qualitatively correct)} \ , \\ \infty & \text{, if } \lambda h > 1 \quad \text{(oscillations, qualitatively wrong)} \ . \end{cases}$$

More extensive discussion of the analysis of the model problem in Sect. 3.1.

Example 1.4.17 (Euler method for pendulum equation).

Mathematical pendulum $\rightarrow$ Ex. 1.2.17: Hamiltonian form (1.2.19) of the equations of movement

$$\text{angular momentum} \quad p := \dot{\alpha} \quad \Rightarrow \quad \frac{d}{dt}\begin{pmatrix} \alpha \\ p \end{pmatrix} = \begin{pmatrix} p \\ -\frac{g}{l}\sin\alpha \end{pmatrix} \ , \quad g = 9.8, l = 1 \ . \qquad (1.2.19)$$

- Approximate numerical solution with explicit/implicit Euler method (1.4.2)/(1.4.13),
- Constant step size $h = T/N$, $T = 5$ final time, $N \in \{50, 100, 200\}$,
- initial values: $\alpha(0) = \pi/4$, $p(0) = 0$.

Listing 1.5: Simulation of the mathematical pendulum using Euler method

```
function pendeul(y0,T,N,filename)
```

P. Grohs

rev 63606, February 17, 2014

```
% MATLAB function applying explicit and implicit Euler methods and implicit
% midpoint rule of Sect. 1.4.3 to mathematical pendulum equation in
% minimal coordinates and Hamiltonian form for Ex. 1.4.17
% Arguments:  y0 ≙ initial position, T ≙ final time N ≙
% number of equidistant timesteps

g = 9.8;  % constant of gravity
l = 1;    % length of pendulum

% Compute 'exact' solution by means of high-order single step method with tight
% error control
odefun = @(t,y) [y(2);-g/l*sin(y(1))];
[t,s] =
  ode45(odefun,[0,T],y0,odeset('abstol',1E-10,'reltol',1E-10));

h = T/N;  % timestep

% Explicit Euler (1.4.2)
y_expl = y0; y = y0;
for k=1:N
  y = y + h*[y(2);-g/l*sin(y(1))];
  y_expl = [y_expl,y];
end
```

```
% Implicit Euler
y_imp = y0; y = y0;
for k=1:N
  % Implicit Euler equation for next angle
  F = @(x) x+h*h*g/l*sin(x) - y(1) - h*y(2);
  [y(1),Fval] = fsolve(F,y(1)+h*y(2)); % solve non-linear system of equations
  fprintf('Impl Euler step %d: residual %f\n',k,Fval);
  y(2) = y(2) - h*g/l*sin(y(1));
  y_imp = [y_imp,y];
end

% Implicit midpoint rule
y_mid = y0; y = y0;
rhs = @(y)  [y(2);-g/l*sin(y(1))];

for k=1:N
  % Implicit equation (1.4.19) for implicit midpoint rule
  F = @(x) (x - h*rhs(y+0.5*x));
  [dy,Fval] = fsolve(F,h*rhs(y)); y = y+dy;
  fprintf('Impl midp step %d: residual %f\n',k,norm(Fval));
  y_mid = [y_mid,y];
end
```

```
tg = h*(0:N);

% Plotting of trajectories in phase space
figure('name','pendeul');
ph = plot(s(:,1),s(:,2),'g--',...
       y_expl(1,:),y_expl(2,:),'r-+',...
       y_imp(1,:),y_imp(2,:),'b-+',...
       y_mid(1,:),y_mid(2,:),'m-*'); hold on;
set(ph(1),'linewidth',2);
ax = axis;
plot([ax(1) ax(2)],[0 0],'k-');
plot([0 0],[ax(3) ax(4)],'k-');
xlabel('{\bf \alpha}','fontsize',14);
ylabel('{\bf p}','fontsize',14);
legend('exact solution','explicit Euler','implicit Euler',...
       'implicit midpoint','location','southwest');
title(sprintf('%d timesteps on [0,%f]',N,T));

if (nargin > 3)
  print('-depsc2',sprintf('%s.eps',filename));
end
```

```
% Tracking energies for explicit Euler
y = y_expl;

E_kin = 0.5*(y(2,:).^2);
E_pot = -g/l*cos(y(1,:));
E_pot = E_pot - min(E_pot) + min(E_kin);
E_tot = E_kin + E_pot;

% Plot of evolution of energies
figure('name','Pendulum: energy');
plot(tg,E_kin,'b-',...
     tg,E_pot,'c-',...
     tg,E_tot,'r-');
xlabel('{\bf time t}','fontsize',14);
ylabel('{\bf energy}','fontsize',14);
legend('kinetic energy','potential energy','total energy');
title('Energies for {\bf explicit} Euler discrete evolution');

if (nargin > 3),
  print('-depsc2',sprintf('%s_EnExpl.eps',filename)); end

% Tracking energies for implicit Euler
y = y_imp;
```

```
E_kin = 0.5*(y(2,:).^2);
E_pot = -g/l*cos(y(1,:));
E_pot = E_pot - min(E_pot) + min(E_kin);
E_tot = E_kin + E_pot;

figure('name','Pendulum: energy');
plot(tg,E_kin,'b-',...
     tg,E_pot,'c-',...
     tg,E_tot,'r-');
xlabel('{\bf time t}','fontsize',14);
ylabel('{\bf energy}','fontsize',14);
legend('kinetic energy','potential energy','total energy');
title('Energies for {\bf implicit} Euler discrete evolution');

if (nargin > 3)
  print('-depsc2',sprintf('%s_EnImpl.eps',filename));
end

% Tracking energies for implicit midpoint rule
y = y_mid;

E_kin = 0.5*(y(2,:).^2);
E_pot = -g/l*cos(y(1,:));
```

```
E_pot = E_pot - min(E_pot) + min(E_kin);
E_tot = E_kin + E_pot;

figure('name','Pendulum: energy');
plot(tg,E_kin,'b-',...
     tg,E_pot,'c-',...
     tg,E_tot,'r-');
xlabel('{\bf time t}','fontsize',14);
ylabel('{\bf energy}','fontsize',14);
legend('kinetic energy','potential energy','total
  energy','location','southwest');
title('Energies for {\bf implicit midpoint} discrete evolution');

if (nargin > 3)
  print('-depsc2',sprintf('%s_EnImid.eps',filename));
end
```

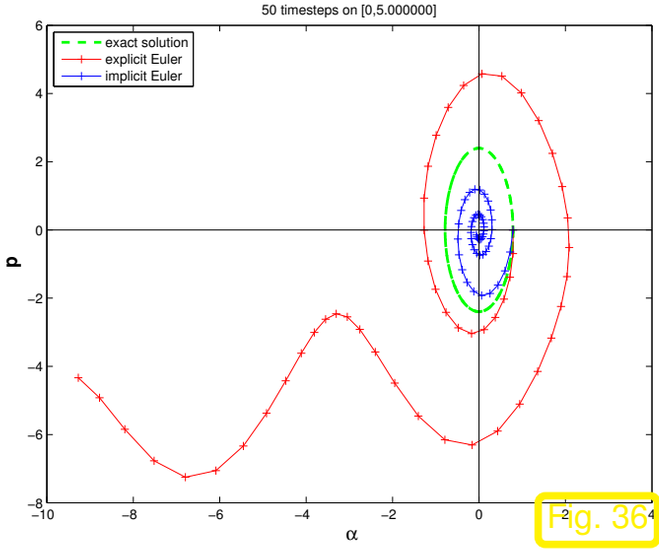


Fig. 36

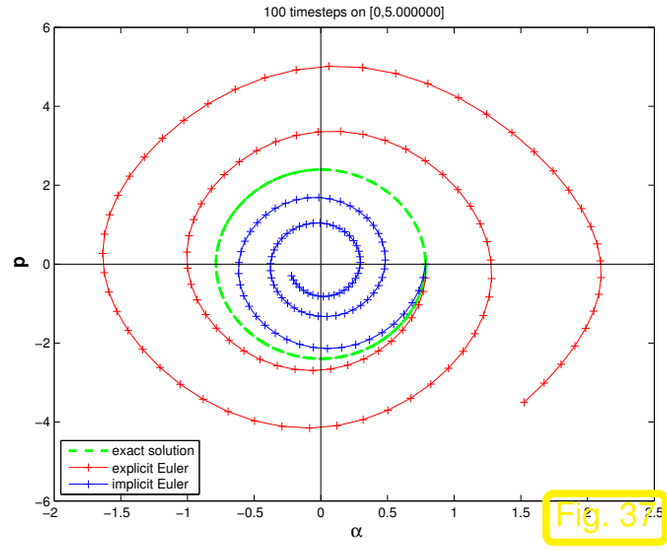


Fig. 37

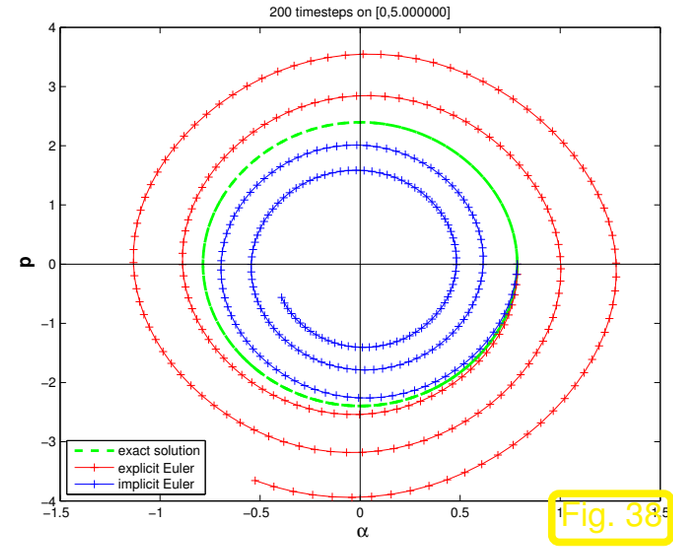


Fig. 38

Behavior of the approximate energies: kinetic energy : $E_{\text{kin}}(t) = \frac{1}{2}p(t)^2$

potential energy : $E_{\text{pot}}(t) = -\frac{g}{l}\cos\alpha(t)$

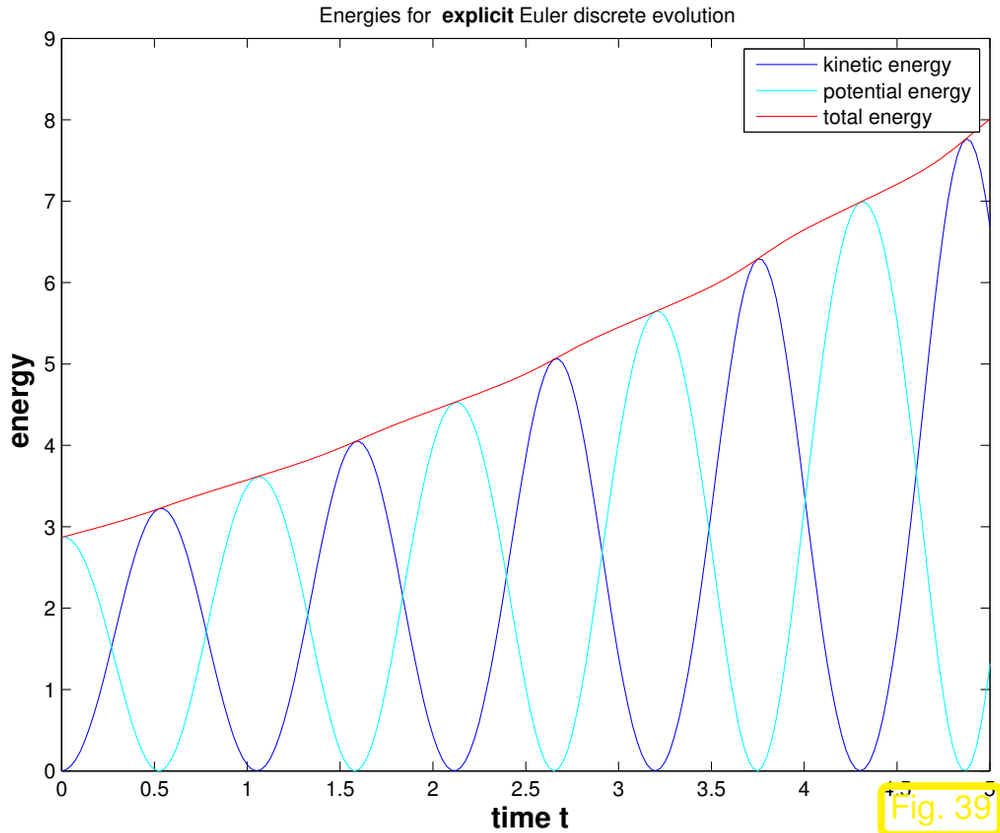


Fig. 39

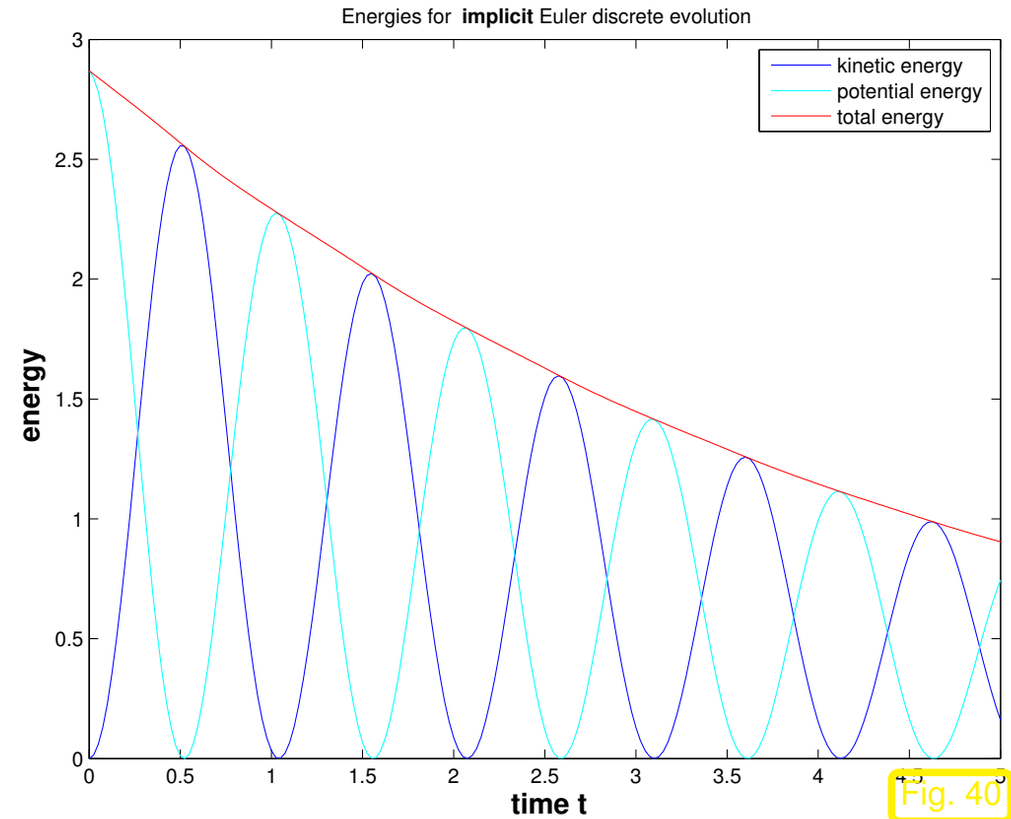


Fig. 40

P. Grohs

rev 63606, February 17, 2014

☞ Explicit Euler: total energy of the pendulum grows

☞ Implicit Euler: total energy of the pendulum decays (“numerical friction”)

◇

Example 1.4.18 (Euler method for length preserving evolution).

Initial value problem for , $D = \mathbb{R}^2$:

$$\dot{\mathbf{y}} = \begin{pmatrix} y_2 \\ -y_1 \end{pmatrix} \quad , \quad \mathbf{y}(0) = \mathbf{y}_0 \quad \blacktriangleright \quad \mathbf{y}(t) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} \mathbf{y}_0 \, .$$

Invariant ($\rightarrow$ Def. 1.2.7): $I(\mathbf{y}) = \|\mathbf{y}\|$

(movement with constant velocity on circular path)

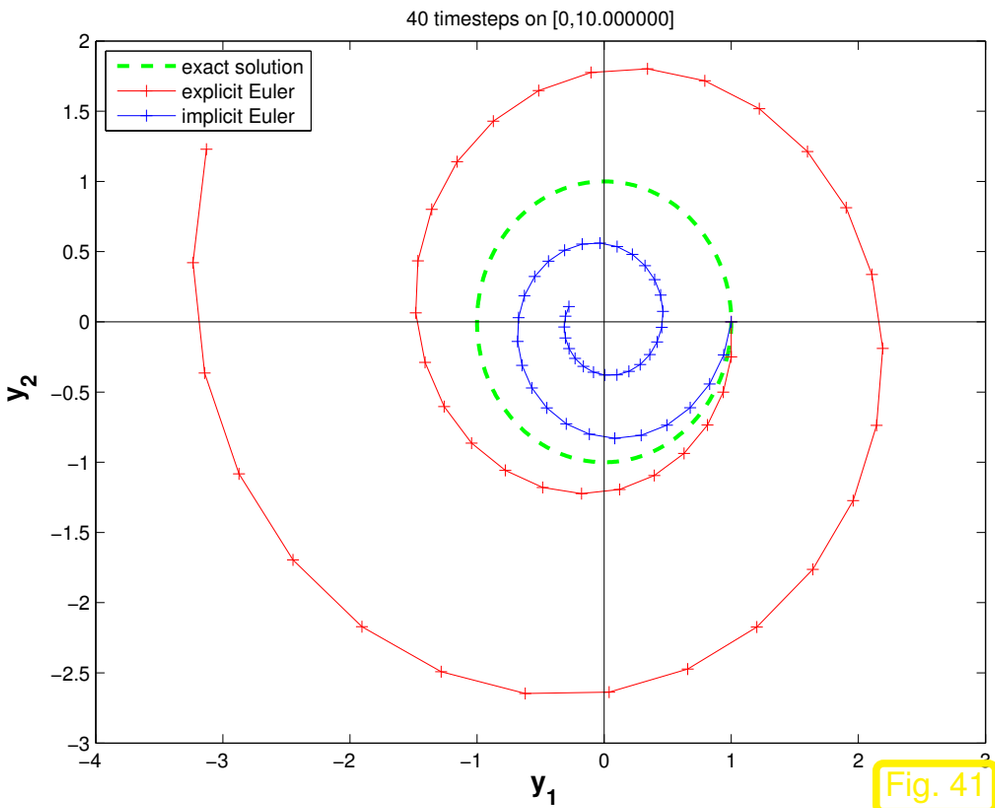


Fig. 41

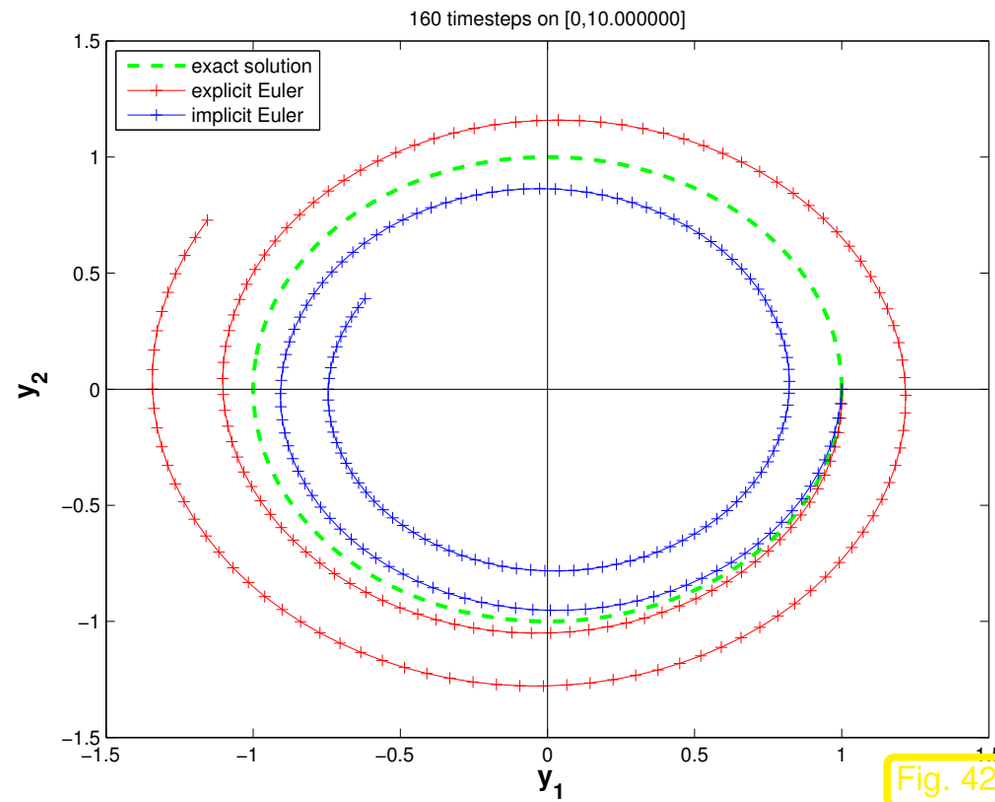


Fig. 42

☞ Explicit Euler: Numerical solution “is carried outward”

☞ Implicit Euler: Numerical solution “collapses into the center”

[Beitrag von J. Mitrovic, FS 2011]

Concluding remark: Both methods discussed so far, the implicit and explicit Euler method, are easy to implement. However, one observes significant decay in the structure of the computed solutions under certain circumstances. For example, energy and length are not preserved for the pendulum equation (despite them being invariants). This is in particular an issue for ODEs derived from physics. Therefore, further methods have been developed, such as the implicit midpoint rule, which is a combination of the two Euler methods and will be presented in the next section.

1.4.3 Implicit midpoint rule

How do we avoid the energy drift for explicit/implicit Euler method applied to conservative systems **?**

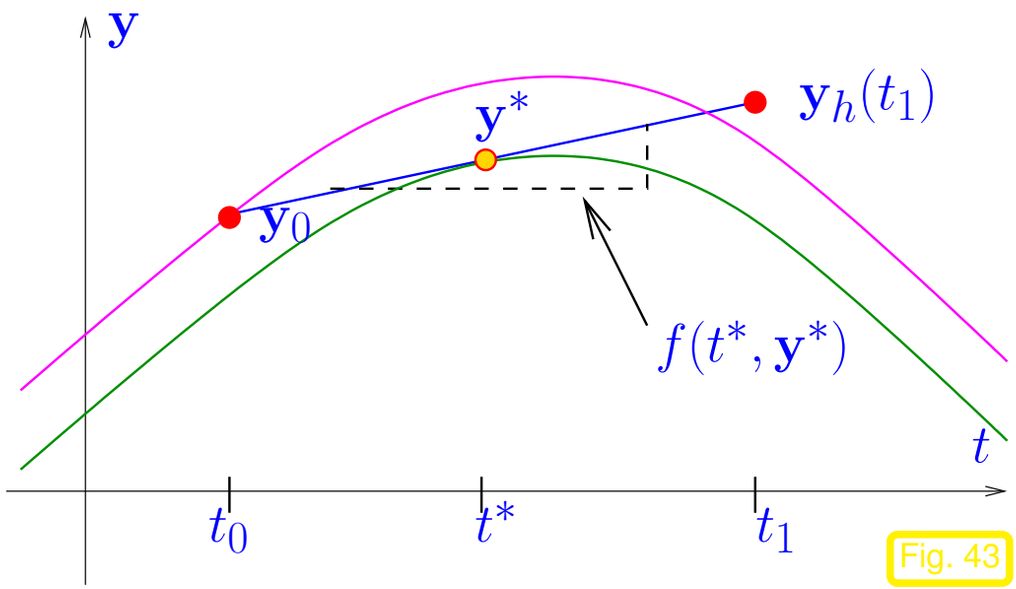


Fig. 43

Idea: approximate solution through $(t_0, \mathbf{y}_0)$ on $[t_0, t_1]$ by

- linear polynomial at $(t_0, \mathbf{y}_0)$
- with slope $f(t^*, \mathbf{y}^*)$, $t^* := \frac{1}{2}(t_0 + t_1)$, $\mathbf{y}^* = \frac{1}{2}(\mathbf{y}_0 + \mathbf{y}_1)$

◁ — $\hat{=}$ solution curve through (t_0, y_0),
— $\hat{=}$ solution curve through (t^*, y^*),
— $\hat{=}$ tangent to — at (t^*, y^*).

Apply to small time intervals $[t_0, t_1], [t_1, t_2], \ldots, [t_{N-1}, t_N]$ ➤ implicit midpoint rule

▶ approximations $\mathbf{y}_{k+1}$ for $\mathbf{y}(t_k)$ generated by the implicit midpoint rule satisfy

$$\boxed{\mathbf{y}_{k+1} := \mathbf{y}_h(t_{k+1}) = \mathbf{y}_k + h_k \mathbf{f}(\tfrac{1}{2}(t_k + t_{k+1}), \tfrac{1}{2}(\mathbf{y}_k + \mathbf{y}_{k+1})) \;, \quad k = 0, \ldots, N-1} \;, \qquad (1.4.19)$$

with local (time) step size $h_k := t_{k+1} - t_k$.

Note: (1.4.19) requires solution of a (potentially non-linear) equation for $\mathbf{y}_{k+1}$!
(➤ terminology “implicit”)

Remark 1.4.20 (Implicit midpoint rule as difference method).

(1.4.19) by approximation of the time derivative $\frac{d}{dt}$ by central difference quotients on time grid $\mathcal{G} := \{t_0, t_1, \dots, t_N\}$:

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}) \quad \longleftrightarrow \quad \frac{\mathbf{y}_h(t_{k+1}) - \mathbf{y}_h(t_k)}{h_k} = \mathbf{f}(\tfrac{1}{2}(t_k + t_{k+1}), \tfrac{1}{2}(\mathbf{y}_h(t_k) + \mathbf{y}_h(t_{k+1})), \; , \qquad k = 0, \dots, N-1 \; .$$

△

Example 1.4.21 (Implicit midpoint rule for logistic differential equation).

Repeat numerical experiments from example 1.4.9 for implicit midpoint rule (1.4.19):

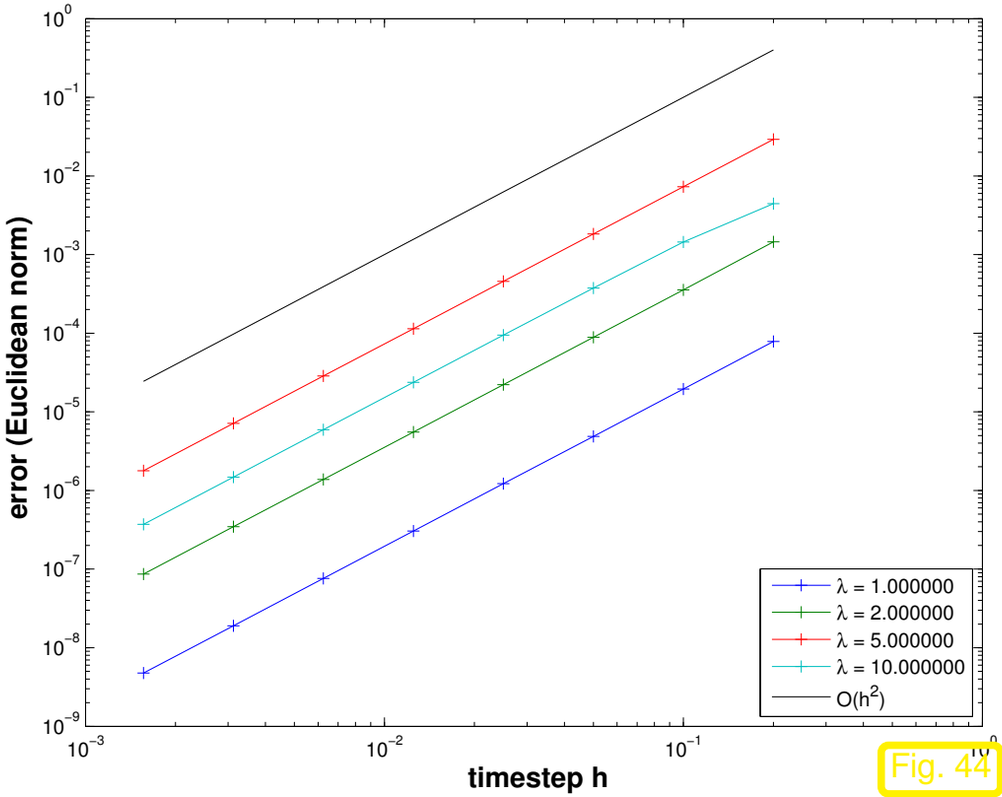


Fig. 44

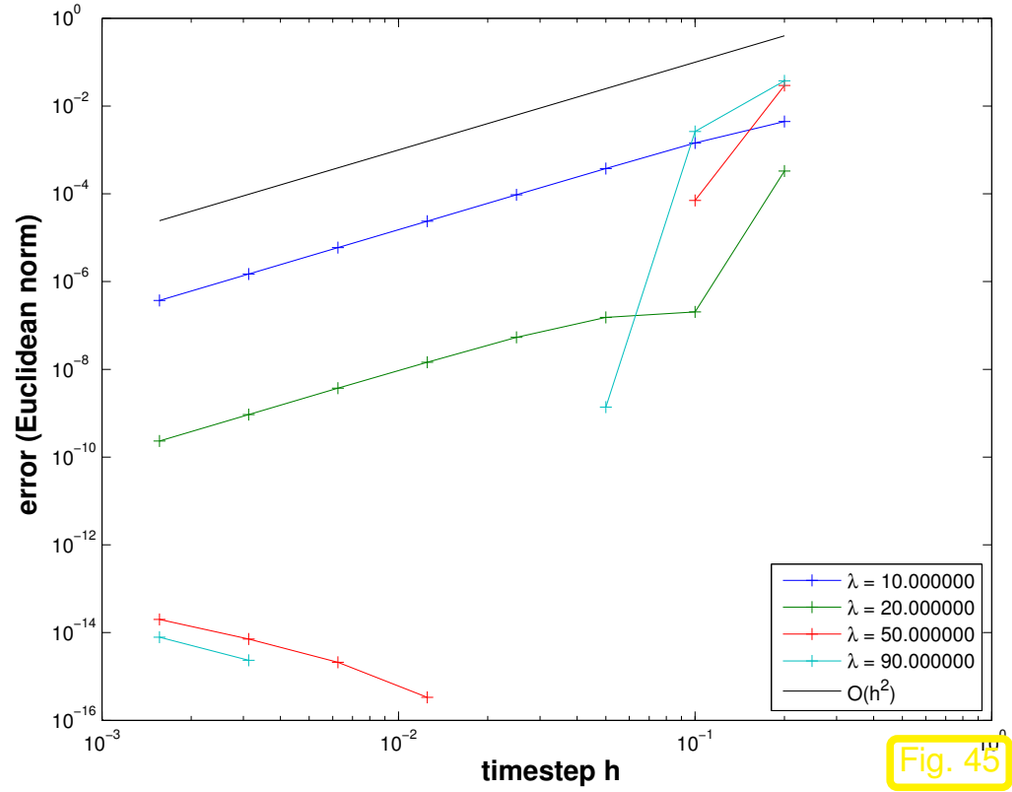


Fig. 45

λ small: $O(h^2)$ convergence (asymptotic)

λ large: stable for all step sizes h **!** ◇

Example 1.4.22 (Implicit midpoint rule for circular movement).

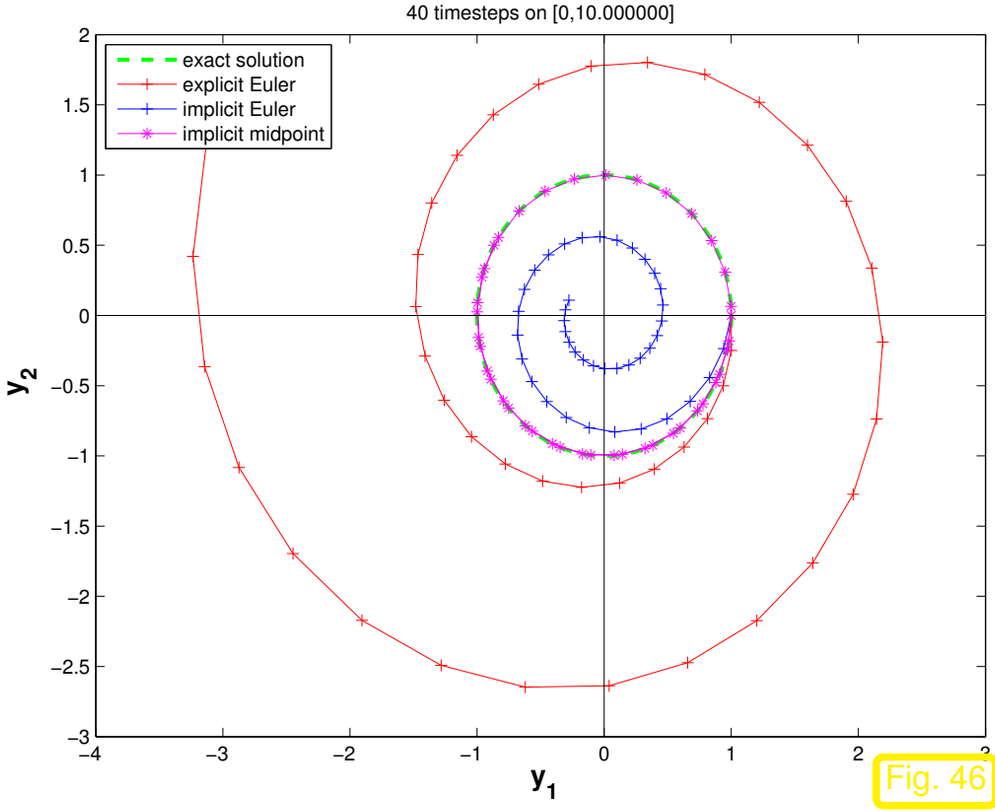


Fig. 46

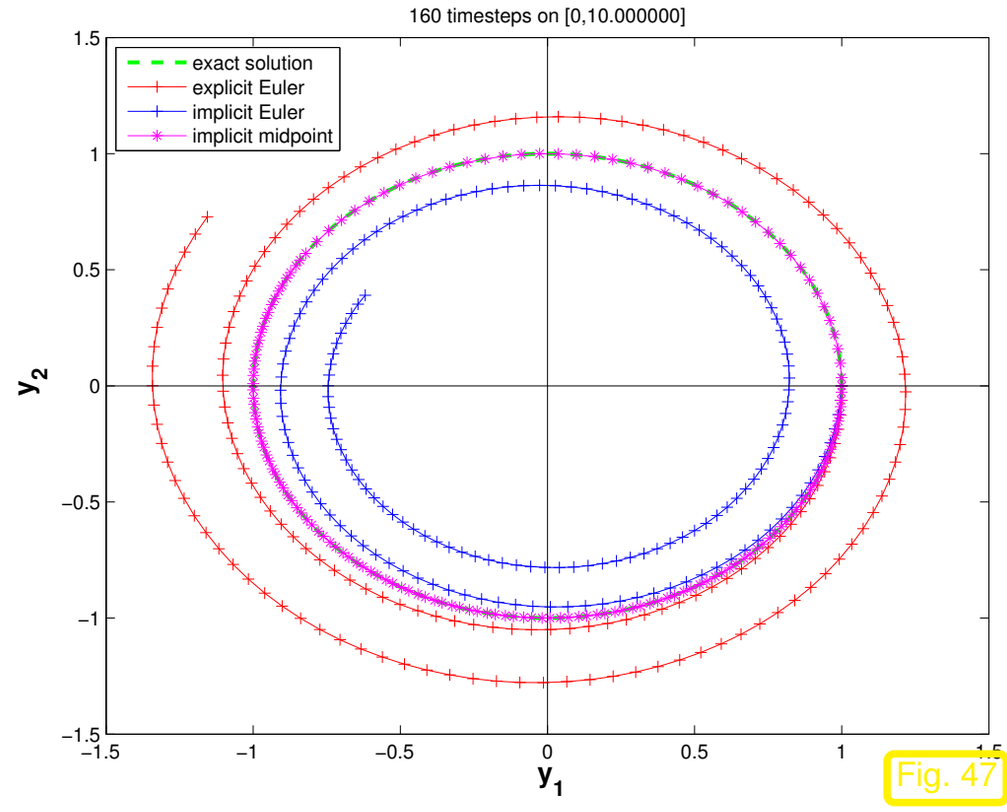


Fig. 47

☞ Implicit midpoint rule: perfect preservation of length **!**

◇

P. Grohs

rev 63606, February 17, 2014

Lemma 1.4.23 (Conservation of quadratic invariants by the implicit midpoint rule)**.**
Let $I : D \subset \mathbb{R}^d \mapsto \mathbb{R}$, $I(\mathbf{y}) := \frac{1}{2}\mathbf{y}^T\mathbf{A}\mathbf{y}$, $\mathbf{A} \in \mathbb{R}^{d,d}$, *be invariant* ($\rightarrow$ *Def. 1.2.7) of the autonomous differential equation* $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ *with globally differentiable right hand side* $\mathbf{f} : D \mapsto \mathbb{R}^d$. *Then, there holds*

$$I(\mathbf{y}_k) = I(\mathbf{y}_0) \quad \forall k \in \mathbb{Z} \quad \textit{for } \mathbf{y}_k \textit{ based on } (1.4.19)$$

[Beitrag von J. Mitrovic, FS 2011]

Proof: We show $I(y_{k+1}) = I(y_k)$ with y_{k+1}, y_k from implicit midpoint rule, for all $k \in \mathbb{N}$. If I invariant of $\dot{\mathbf{y}} = f(t,\mathbf{y}) \Rightarrow f(t,\mathbf{y}(t))^T \cdot \mathbf{grad}(I(\mathbf{y}(t))) = 0$. Chain rule implies that $\mathbf{grad}\, I(\mathbf{y}(t)) = \mathbf{grad}\, \frac{1}{2}\mathbf{y}^T A\mathbf{y} = \mathbf{grad}\, \frac{1}{2}\sum_{i,j=1}^n y_i a_{ij} y_j = A\mathbf{y}$, as A symmetric. $\Rightarrow f(t,\mathbf{y}(t))^T A\mathbf{y} = 0 \quad \forall \mathbf{y} \in D$ $I(y_{k+1}) - I(y_k) = \frac{1}{2}y_{k+1}^T A y_{k+1} - \frac{1}{2}y_k^T A y_k = \frac{1}{2}(y_{k+1} - y_k)^T A(y_{k+1} + y_k) = hf(\frac{1}{2}(t_{k+1} + t_k), \frac{1}{2}(y_{k+1} + y_k))^T A(\frac{1}{2}(y_{k+1} + y_k)) = 0$, as I is an invariant. This implies that the invariant is preserved.

Lemma 1.4.23 represents a special case of a structural property that is preserved by a specific class of implicit one step methods. The most simple representative of this class is the implicit midpoint rule, compare Theorem 4.1.4.

Example 1.4.24 (Implicit midpoint rule for pendulum equation).

Initial value problem and numerical experiments as in Ex. 1.4.17

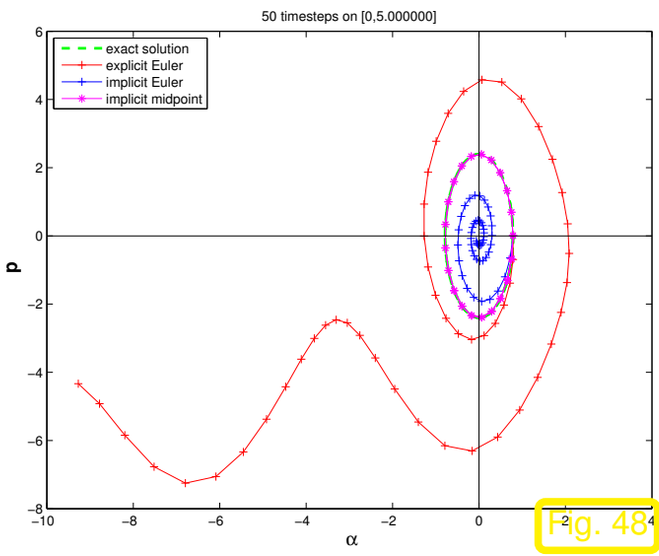


Fig. 48

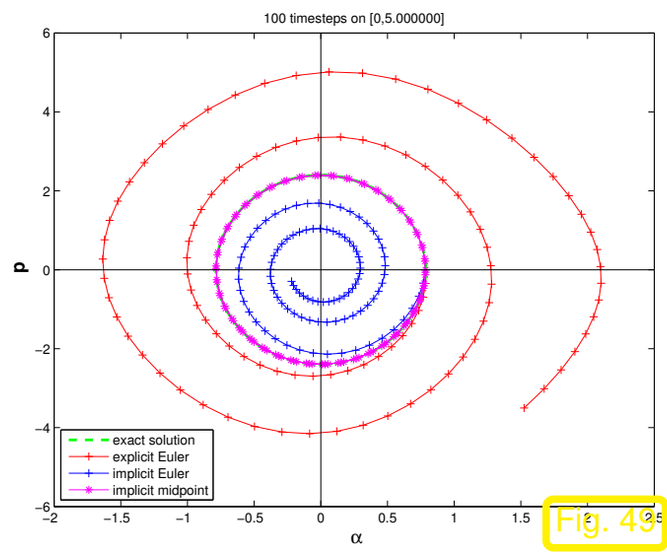


Fig. 49

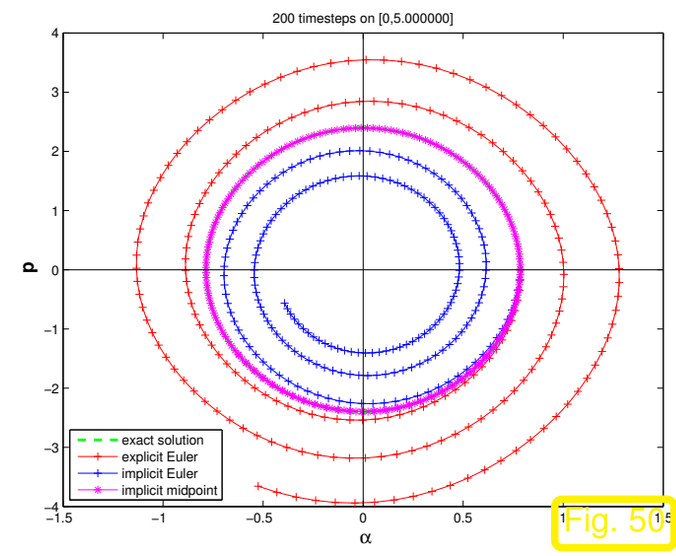


Fig. 50

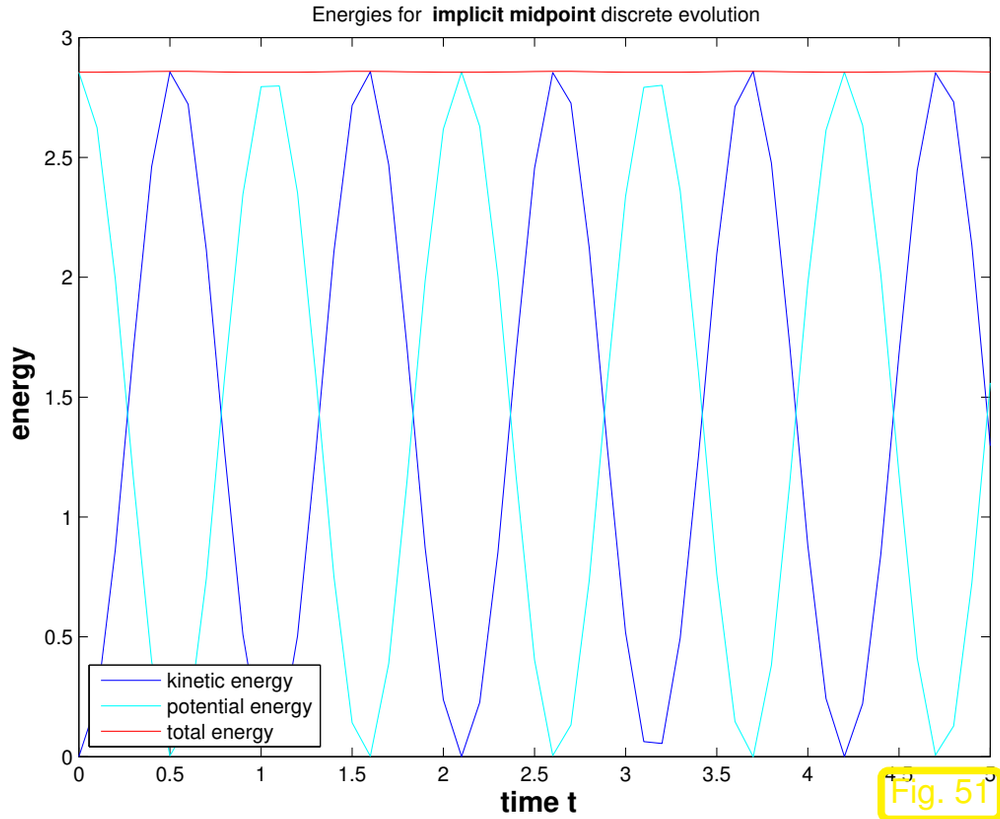


Fig. 51

◁ Behavior of energies for numerical integration with implicit midpoint rule (1.4.19), $N = 50$.

No (visual) energy drift when compared with the Euler method (despite large time steps)

◇

P. Grohs

rev 63606, February 17, 2014

[Beitrag von J. Mitrovic, FS 2011]

We remember the following for the implicit midpoint rule (1.4.19):

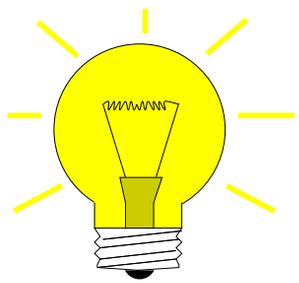
- Solutions generated by the implicit midpoint rule converge algebraically with order 2, i.e., the error behaves like $O(h^2)$.
- The midpoint rule preserves quadratic integrals exactly such as lengths and the energy for equations of movement taken as time average.

1.4.4 Störmer-Verlet method [20]

Transfer of the idea of the Euler methods (→ Sect. 1.4.1, 1.4.2) to second-order differential equations

?

$$\ddot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}) \ . \tag{1.4.25}$$



Given $\mathbf{y}_{k-1} \approx \mathbf{y}(t_{k-1})$, $\mathbf{y}_k \approx \mathbf{y}(t_k)$ approximate $\mathbf{y}(t)$ on $[t_{k-1}, t_{k+1}]$ by

- Parabola $\mathbf{p}(t)$ through $(t_{k-1}, \mathbf{y}_{k-1})$, $(t_k, \mathbf{y}_k)$ $(*)$,
- with $\ddot{\mathbf{p}}(t_k) = \mathbf{f}(\mathbf{y}_k)$ $(*)$.

$(*)$ ➜ Parabola uniquely determined.

$$\mathbf{y}_{k+1} := \mathbf{p}(t_{k+1}) \approx \mathbf{y}(t_{k+1})$$

Störmer-Verlet method for (1.4.25) (time grid $\mathcal{G} := \{t_0, t_1, \dots, t_N\}$):

$$\mathbf{y}_{k+1} = -\frac{h_k}{h_{k-1}}\mathbf{y}_{k-1} + \left(1 + \frac{h_k}{h_{k-1}}\right)\mathbf{y}_k + \tfrac{1}{2}(h_k^2 + h_k h_{k-1})\mathbf{f}(t_k, \mathbf{y}_k) \ , \quad k = 1, \dots, N-1 \ . \tag{1.4.26}$$

for constant time step sizes h:

$$\mathbf{y}_{k+1} = -\mathbf{y}_{k-1} + 2\mathbf{y}_k + h^2\mathbf{f}(t_k, \mathbf{y}_k) \;, \quad k = 1, \dots, N-1 \;. \tag{1.4.27}$$

Note: (1.4.26) does not require the solution of an equation (➤ explicit method)

Terminology: $\mathbf{y}_{k+1} = \mathbf{y}_{k+1}(\mathbf{y}_k, \mathbf{y}_{k-1})$ ➤ (1.4.26) is a two step method
(Explicit/implicit Euler method, midpoint rule = one/single step methods)

Remark 1.4.28 (Störmer-Verlet method as difference method).

(1.4.27) by approximating the second derivatives w.r.t. time by second-order central difference quotients on time grid $\mathcal{G} := \{t_0, t_1, \dots, t_N\}$: for uniform step size $h > 0$

$$\ddot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) \quad \longleftrightarrow \quad \frac{\frac{\mathbf{y}_h(t_{k+1}) - \mathbf{y}_h(t_k)}{h} - \frac{\mathbf{y}_h(t_k) - \mathbf{y}_h(t_{k-1})}{h}}{h} = \frac{\mathbf{y}_h(t_{k+1}) - 2\mathbf{y}_h(t_k) + \mathbf{y}_h(t_{k-1})}{h^2} = \mathbf{f}(\mathbf{y}_h(t_k)) \;.$$

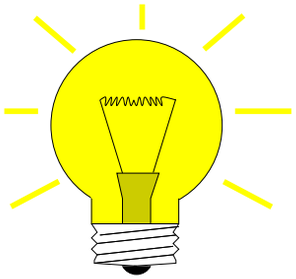
△

P. Grohs

rev 63606, February 17, 2014

Remark 1.4.29 (Initial step for Störmer-Verlet method).

initial values for (1.4.25), compare Rem. 1.1.16: $\mathbf{y}(0) = \mathbf{y}_0$, $\dot{\mathbf{y}}(0) = \mathbf{v}_0$



- Use virtual point in time $t_{-1} := t_0 - h_0$
- Apply (1.4.27) on $[t_{-1}, t_1]$:

$$\mathbf{y}_1 = -\mathbf{y}_{-1} + 2\mathbf{y}_0 + h_0^2 \mathbf{f}(t_0, \mathbf{y}_0) \ . \tag{1.4.30}$$

- Central difference quotients on $[t_{-1}, t_1]$:

$$\frac{\mathbf{y}_1 - \mathbf{y}_{-1}}{2h_0} = \mathbf{v}_0 \ . \tag{1.4.31}$$

Compute $\mathbf{y}_1$ from (1.4.30) **&** (1.4.31) △

Example 1.4.32 (Störmer-Verlet method for pendulum equation).

Listing 1.6: Störmer-Verlet method for pendulum equation

```
function sverletpend(y0,v0,T,N,filename)
% MATLAB function for Stoermer-Verlet simulation of movement of mathematical
% pendulum, Example 1.4.32
%
% y0,v0:  initial values for angle and its temporal derivative
% T : final time
% N : number of timesteps

g = 9.8; l = 1;

% right hand side for (1.4.25) (Newton's equations of motion)
f = @(y) -g/l*sin(y);

h = T/N; % uniform timestep
y_old = y0; y_new = h*v0+y0+ 0.5*h*h*f(y0); % initial step, see
  Rem. 1.4.29

% Stoermer-Verlet iteration
y_sv = y0; y_sv = [y_sv,y_new]; y_p = v0;
for k=2:N+1
  y = -y_old + 2*y_new + h*h*f(y_new);
  y_p = [y_p,(y-y_old)/(2*h)];
  y_old = y_new; y_new = y;
```

```
  y_sv = [y_sv,y];
end
y_sv = y_sv(1:end-1);

% right hand side (Hamiltonian form) for computation of reference solution
% with high-order single step method with tight tolerances
odefun = @(t,y) [y(2);-g/l*sin(y(1))];
[t,y] = ode45(odefun,[0,T],[y0;v0],...
              odeset('abstol',1E-11,'reltol',1E-11,'stats','on'));

% Plot of angle vs. time
figure('name','Pendulum alpha');
plot(t,y(:,1),'g-',h*(0:N),y_sv,'r-+');
xlabel('{\bf time t}','fontsize',14);
ylabel('{\bf angle \alpha}','fontsize',14);
title(sprintf('Pendulum g = %f, l =
  %f,\\alpha(0)=%f,p(0)=%f',g,l,y0,v0));
if (nargin > 3),
  print('-depsc2',sprintf('%s_alpha.eps',filename)); end

% Plot of velocity vs. time
tg = h*(0:N);
figure('name','Pendulum velocity');
```

```
plot(t,y(:,2),'g-',tg,y_p,'r-+');
xlabel('{\bf time t}','fontsize',14);
ylabel('{\bf velocity p}','fontsize',14);
title(sprintf('Pendulum g = %f, l = %f,\\alpha(0)=%f,p(0)=%f',g,l,y0,v0));
if (nargin > 3), print('-depsc2',sprintf('%s_p.eps',filename)); end

% PLot of trajectory in phase space
figure('name','Pendulum trajectory');
ph = plot(y(:,1),y(:,2),'g--',y_sv,y_p,'r-+');
set(ph(1),'linewidth',2);
xlabel('{\bf angle \alpha}','fontsize',14);
ylabel('{\bf velocity p}','fontsize',14);
title(sprintf('Pendulum g = %f, l = %f,\\alpha(0)=%f,p(0)=%f',g,l,y0,v0));

if (nargin > 3), print('-depsc2',sprintf('%s_orbit.eps',filename)); end

% Tracking of energies
E_kin = 0.5*(y_p.^2);
E_pot = -g/l*cos(y_sv);
```

```
E_pot = E_pot - min(E_pot) + min(E_kin);
E_tot = E_kin + E_pot;

figure('name','Pendulum: energy');
plot(tg,E_kin,'b-',tg,E_pot,'c-',tg,E_tot,'r-');
xlabel('{\bf time t}','fontsize',14);
ylabel('{\bf energy}','fontsize',14);
legend('kinetic energy','potential energy','total
  energy','location','southeast');
title('Energies for {\bf Stoermer-Verlet} discrete evolution');
if (nargin > 3),
  print('-depsc2',sprintf('%s_EnSV.eps',filename)); end
```

- (1.4.27) applied to (1.2.18)
- Initial step according to Rem. 1.4.29
- Constant step size $h := T/N$, $N \in \mathbb{N}$ time steps
- Reference solution by MATLAB function ode45() (very small tolerances)
- $\alpha_0 = \pi/2$, $p_0 = 0$, $T = 5$, cmp. Ex. 1.4.17
- Number time steps: $N = 40$

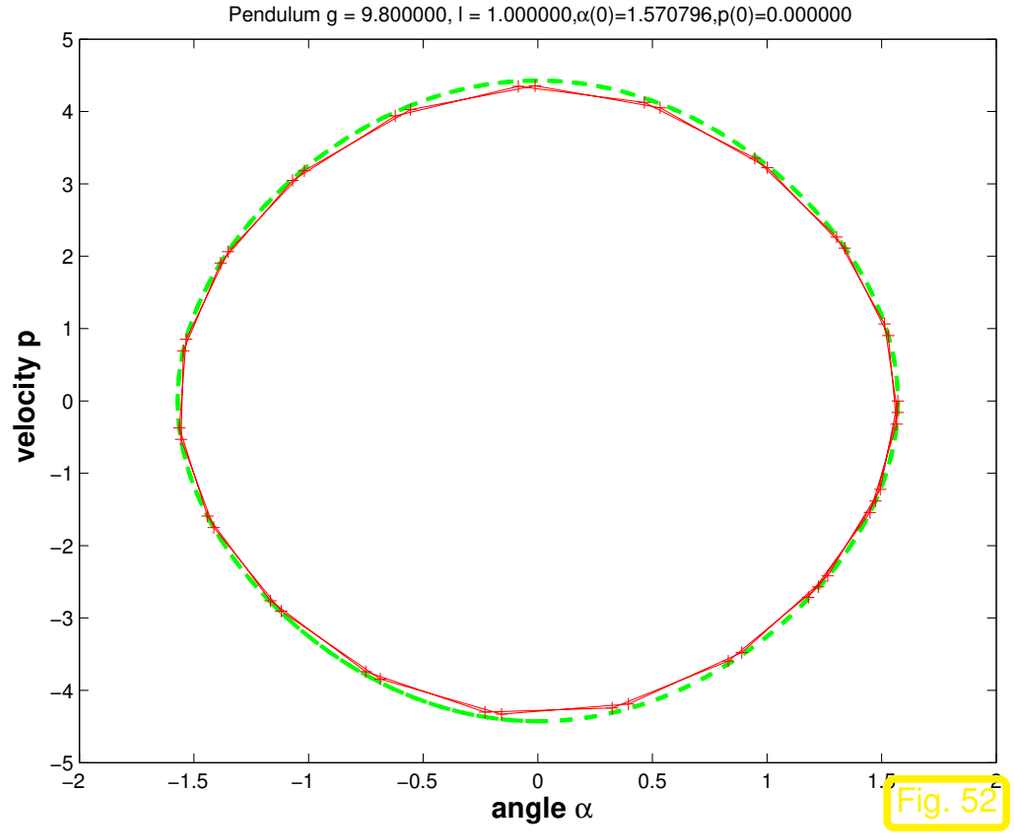


Fig. 52

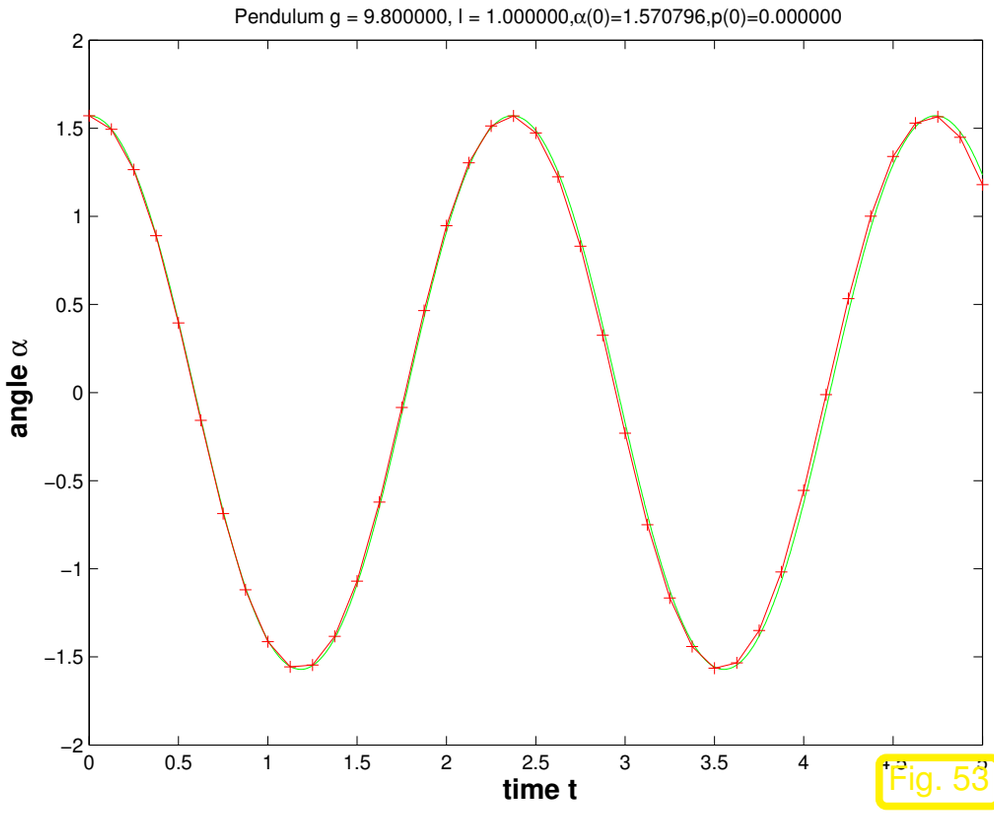


Fig. 53

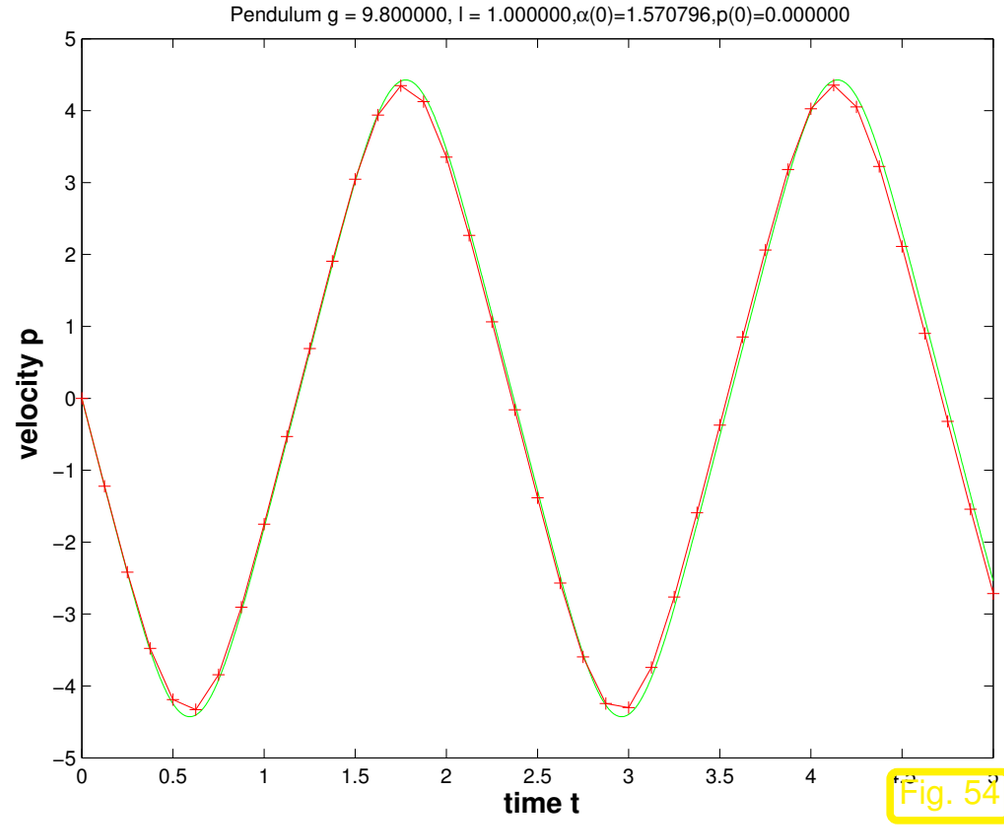


Fig. 54

P. Grohs

rev 63606, February 17, 2014

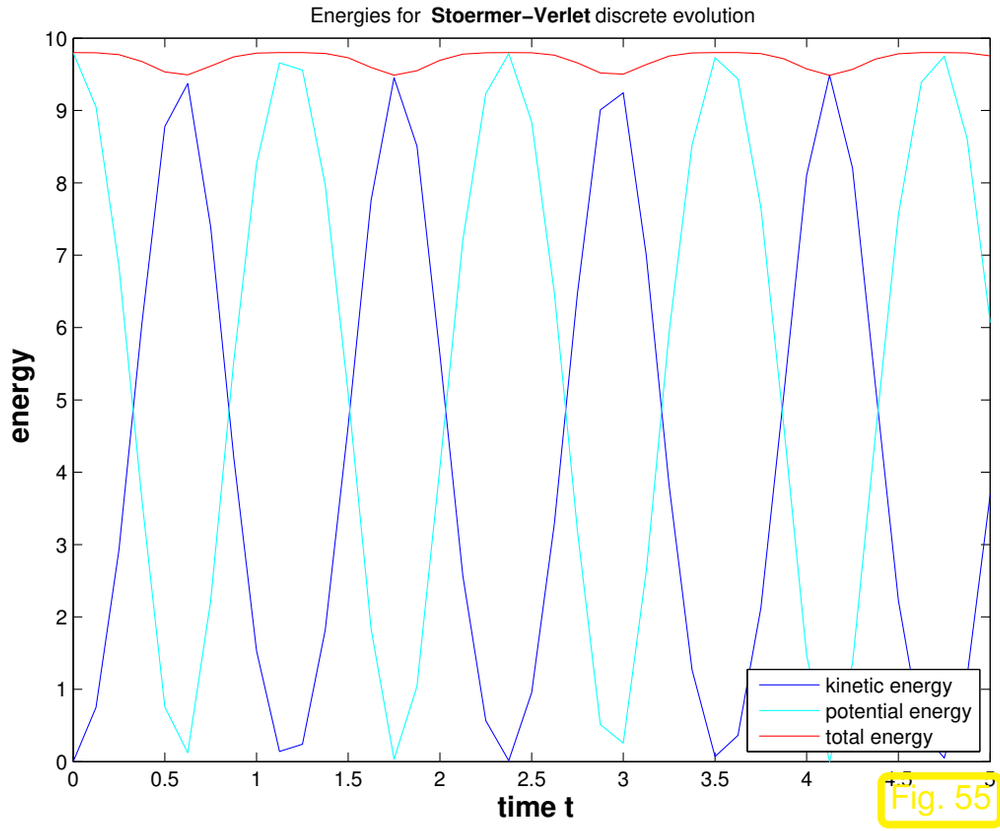


Fig. 55

☞ No energy drift despite large time steps

Perfect periodic orbits **!**

Contrast with: Ex. 1.4.17

◇

P. Grohs

rev 63606, February 17, 2014

Remark 1.4.33 (Single step formulation of Störmer-Verlet method)*.*

for constant time steps, cmp. (1.4.27), analogously to transformation of a second-order differential

equation to a first-order differential equation, compare (1.1.10): with $\mathbf{v}_{k+\frac{1}{2}} := \frac{\mathbf{y}_{k+1}-\mathbf{y}_k}{h}$

$$\begin{array}{ccc} \ddot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) & \longleftrightarrow & \begin{aligned} \dot{\mathbf{y}} &= \mathbf{v} \,, \\ \dot{\mathbf{v}} &= \mathbf{f}(\mathbf{y}) \,. \end{aligned} \\ \updownarrow & & \updownarrow \\ \mathbf{y}_{k+1} - 2\mathbf{y}_k + \mathbf{y}_{k-1} = h_k^2 \mathbf{f}(\mathbf{y}_k) & \longleftrightarrow & \boxed{\begin{aligned} \mathbf{v}_{k+\frac{1}{2}} &= \mathbf{v}_k + \tfrac{h}{2}\mathbf{f}(\mathbf{y}_k) \,, \\ \mathbf{y}_{k+1} &= \mathbf{y}_k + h\mathbf{v}_{k+\frac{1}{2}} \,, \\ \mathbf{v}_{k+1} &= \mathbf{v}_{k+\frac{1}{2}} + \tfrac{h}{2}\mathbf{f}(\mathbf{y}_{k+1}) \,. \end{aligned}} \\ \text{two step method} & & \text{single step method} \end{array}$$

Initial step ($\rightarrow$ Rem. 1.4.29) is implicitly contained in the single step formulation.

$\triangle$

[Beitrag von J. Mitrovic, FS 2011]

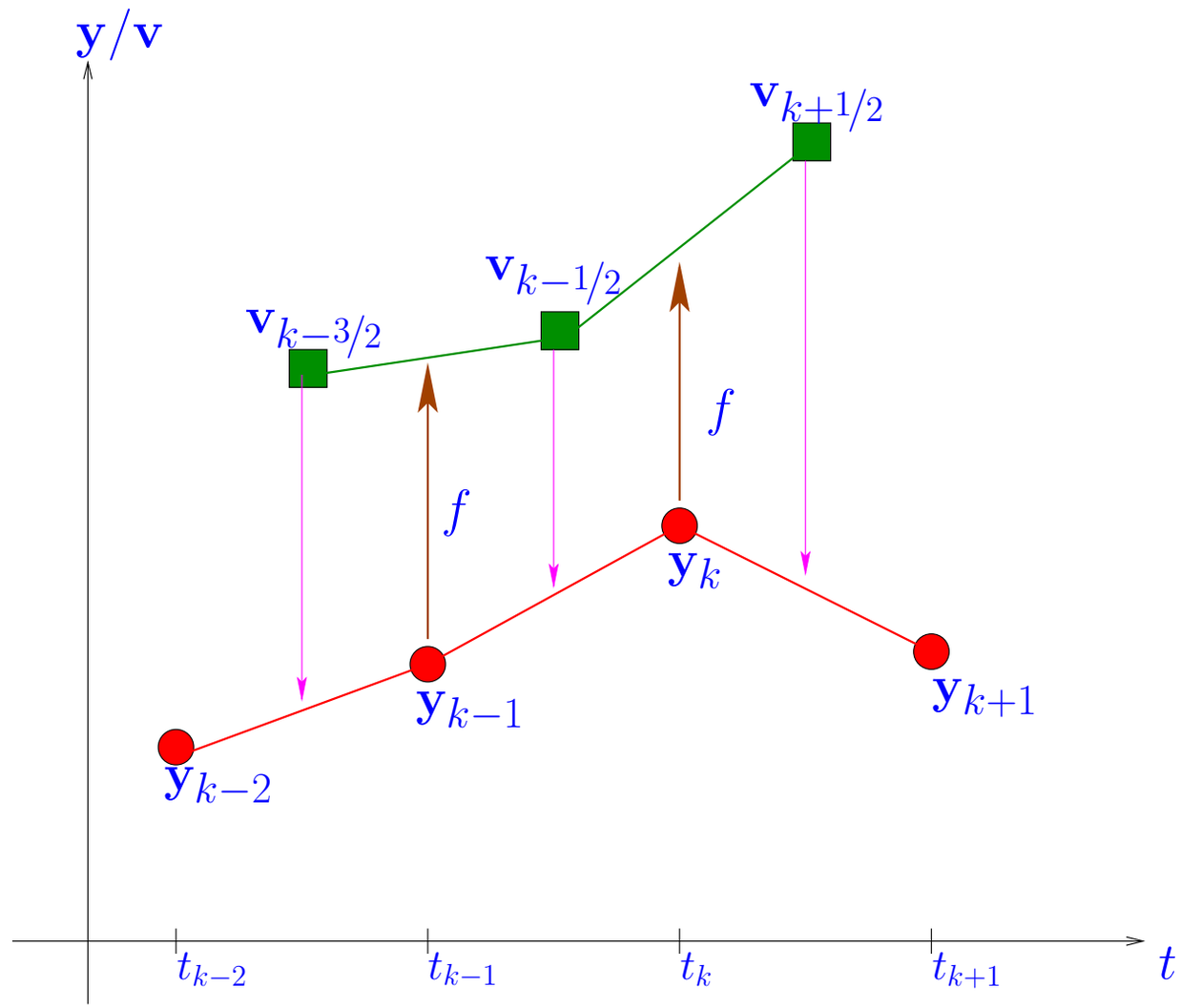
The single step formulation of the Störmer-Verlet method is an explicit method and therefore cheap (,as one does not need to solve a system of equations). Also, it preserves the energy almost perfectly

(the energy does not drift, it just oscillates a bit). Mostly for this reason, this method is used quite often for the simulation of conservative, mechanical systems. More explanations can be found in Section 4.4.

Remark 1.4.34 (Störmer-Verlet method as polygonal line method).

Perspective: Störmer-Verlet method as single step method (compare Rem. 1.4.33)

$$\begin{aligned} \mathbf{v}_{k+\frac{1}{2}} &= \mathbf{v}_{k-\frac{1}{2}} + h\mathbf{f}(\mathbf{y}_k) \ , \\ \mathbf{y}_{k+1} &= \mathbf{y}_k + h\mathbf{v}_{k+\frac{1}{2}} \ . \end{aligned}$$



Remember (Rem. 1.2.4) the question "Why are there that many different methods to numerically solve ODEs?"

Answer: Every numerical solver/integrator has special properties

➥ particularly useful/unsuitable for certain classes of IVPs

2 Single step methods

2.1 Basics

Given: $\mathbf{f} : \Omega \mapsto \mathbb{R}^d$ locally Lipschitz continuous ($\rightarrow$ Def. 1.3.2) on augmented state space $\Omega \subset \mathbb{R} \times D$

➢ Defines ODE $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$ ($\rightarrow$ Sect. 1.1)

► Corresponding evolution: $\mathbf{\Phi}^{s,t} : D \mapsto D$ ($\rightarrow$ Def. 1.3.7)

Given: initial values $(t_0, \mathbf{y}_0) \in \Omega$ ➢ specific initial value problem (1.1.13)

Goal: ☞ Approximation of $\mathbf{y}(T)$ for final time $T \in J(t_0, y_0)$.
☞ Approximation of function $t \mapsto \mathbf{y}(t)$, $t \in [t_0, T]$, $T \in J(t_0, \mathbf{y}_0)$ ▷ $\mathbf{y}_h(t)$.

P. Grohs

rev 63606, February 17, 2014

Remark 2.1.1 (Smoothness assumption for right hand side $\mathbf{f}$).

For the convergence theory of single step methods:

Assumption: $\mathbf{f}$ "sufficiently" smooth $\Rightarrow$ $t \mapsto \mathbf{y}(t)$ "sufficiently" smooth

- Initially, proofs are deduced for arbitrarily smooth $\mathbf{f}$.
- Later on, we specify minimal requirements for the smoothness of $\mathbf{f}$ for the statements to be true. (This part is typically not covered in this course.)

△

2.1.1 Abstract single step methods [12, Sect. 4.1]

- Building block: discrete evolution $\boxed{\mathbf{\Psi} : \widetilde{\Omega}_h \subset I \times I \times D \mapsto \mathbb{R}^d}$

  ✎ Notation: $\mathbf{\Psi}^{s,t}\mathbf{y} := \mathbf{\Psi}(s,t;\mathbf{y})$

- Building block: time grid $\boxed{\mathcal{G} := \{t_0, t_1, \dots, t_N = T\}}$, $t_0 < t_1 < \cdots < t_N$.

  (Terminology: $t_k \hat{=}$ grid points, local (time) step size $h_k := t_{k+1} - t_k$)

  ✎ Notation: global time step size $h = h_{\mathcal{G}} = \max\limits_{0 \le i < N} h_k$

Definition 2.1.2 (Single step method)**.**
Given: discrete evolution $\mathbf{\Psi}$ *and time grid* $\mathcal{G} := \{t_0 < t_1 < \cdots < t_N = T\}$*. The recursion*

$$\mathbf{y}_{k+1} := \mathbf{\Psi}(t_k, t_{k+1}; \mathbf{y}_k) \;, \quad k = 0, \dots, N-1 \;, \tag{2.1.3}$$

defines a single step method (or one step method*) for the initial value problem* (1.1.13)*.*

P. Grohs

rev 63606, February 17, 2014

Remark 2.1.4 (Notation for single step method). Often, one specifies single step methods by its first step

$$\mathbf{y}_1 = \text{expression in } \mathbf{y}_0 \text{ and } \mathbf{f} \ .$$

We will sometimes follow this habit in this course.

△

single step method

```
function [T,Y] = esv(Psi,tspan,y0)
t = tspan(1); y = y0; Y = y; T = t;
while (t < tspan(2))
  h = actual time step size
  y = Psi(t,t+h,y); t = t+h;
  Y = [Y,y]; T = [T,t];
end
```

Function handle

`Psi = @(t0,t1,y) ...`

Note: The actual step size is computed based on accuracy requirements and $\mathbf{y}_k$ ($\rightarrow$ Sect. 2.6).

Definition 2.1.5 (Explicit and implicit single step method)**.**
A single step method to approximately solve an IVP is called explicit, if the underlying discrete evolution can be realized by finitely many f-evaluations.

The discrete evolution of an implicit single step method requires to solve a system of equations.

SSM + initial value + time grid generates grid function $\mathbf{y}_{\mathcal{G}} : \mathcal{G} \mapsto \mathbb{R}^d$, $\mathbf{y}_{\mathcal{G}}(t_k) = \mathbf{y}_k$

For a "suitable choice" of $\mathbf{\Psi}$: $\mathbf{y}_k \approx \mathbf{y}(t_k)$ ($\mathbf{y} \hat{=}$ exact solution)

Definition 2.1.6 (Discretization error)**.**

- *For given* $T \in J(t_0, \mathbf{y}_0)$*, let* $\mathbf{y} : [t_0, T] \mapsto \mathbb{R}^d$ *be the solution of the IVP* (1.1.13)
- *Let* $\mathbf{y}_{\mathcal{G}}$ *be an approximate solution on the time grid* $\mathcal{G} = \{t_0 < t_1 < \cdots < t_N = T\}$*.*

discretization error $\epsilon_{\mathcal{G}} := \max\limits_{0 \le k \le N} \|\mathbf{y}(t_k) - \mathbf{y}_k\|$.

P. Grohs

rev 63606, February 17, 2014

Here, $\|\cdot\|$ is an arbitrary vector norm on the state space $D \subset \mathbb{R}^d$. Due to the equivalence of all norms on finite dimensional spaces, all statements that will be deduced in the following are true for arbitrary norms.

Definition 2.1.7 (Convergence and order of convergence). $\rightarrow$ *[12, Def. 4.6]*

The SSM (2.1.3) *for the IVP* (1.1.13) *converges, if*

$$\forall \epsilon > 0: \quad \exists \delta > 0: \quad \forall \text{time grid } \mathcal{G} \subset [0,T]: \quad h_{\mathcal{G}} \leq \delta \quad \Rightarrow \quad \begin{array}{l} \text{SSM well-defined} \\ \epsilon_{\mathcal{G}} \leq \epsilon \ . \end{array}$$

(Short: $\epsilon_{\mathcal{G}} \rightarrow 0$*, if* $h_{\mathcal{G}} \rightarrow 0$*)*

The SSM is called (algebraically $\rightarrow$ *Def. 1.4.5) convergent with order* $p \in \mathbb{N}$*, if*

$$\exists h_0 > 0, \ C > 0: \quad \begin{array}{c} \text{SSM well-defined,} \\ \epsilon_{\mathcal{G}} \leq C h_{\mathcal{G}}^p \end{array} \quad \forall \text{time grid } \mathcal{G}, \ h_{\mathcal{G}} \leq h_0 \ .$$

(Short notation with Landau symbol $\epsilon_{\mathcal{G}} = O(h^p)$*)*

Extension: convergence for all $(t_0, \mathbf{y}_0) \in \Omega$ ➢ global convergence

Note: Convergence according to Def. 2.1.7 is an asymptotic concept ($h_{\mathcal{G}} \rightarrow 0$)

The statement that a method converges with a specific order, typically *does not* imply information about the actual size of the error (measured in a given norm). Such stronger statements are typically not found in the numerical analysis of single step methods.

Why is it useful to know the order of convergence p of a method **?**

Assuming the statement is sharp, i.e., $\epsilon_{\mathcal{G}} \approx Ch_{\mathcal{G}}^p$, we can deduce by which factor we need to reduce the global time step size in order to reduce the error by a given factor.

2.1.2 Consistency [12, Sect. 4.1.1]

Continuous evolution ($\rightarrow$ Def. 1.3.7) $\longleftrightarrow$ discrete evolution

$$\mathbf{\Phi}^{s,t} \qquad\qquad \mathbf{\Psi}^{s,t}$$

satisfies for all $(t,\mathbf{y}) \in \Omega$

(i) $\mathbf{\Phi}^{t,t}\mathbf{y} = \mathbf{y}$

(ii) $\left.\frac{d}{ds}\mathbf{\Phi}^{t,t+s}\mathbf{y}\right|_{s=0} = \mathbf{f}(t,\mathbf{y})$

(iii) $\mathbf{\Phi}^{r,s}\mathbf{\Phi}^{t,r}\mathbf{y} = \mathbf{\Phi}^{t,s}\mathbf{y}\ \forall r,s \in J(t,\mathbf{y})$

should satisfy:

(i) $\mathbf{\Psi}^{t,t}\mathbf{y} = \mathbf{y}$ sure!

(ii) $\left.\frac{d}{ds}\mathbf{\Psi}^{t,t+s}\mathbf{y}\right|_{s=0} = \mathbf{f}(t,\mathbf{y})$ definitely!

(iii) $\mathbf{\Psi}^{r,s}\mathbf{\Psi}^{t,r}\mathbf{y} = \mathbf{\Psi}^{t,s}\mathbf{y}\ \forall r,s \in J(t,\mathbf{y})$ impossible!

Assuming that $t \mapsto \mathbf{\Psi}^{s,t}\mathbf{y}$ is differentiable:

If $\mathbf{\Psi}$ satisfies (i)–(iii), then there holds $\mathbf{\Psi} = \mathbf{\Phi}$ **!**

$$\frac{d}{dt}\left(\mathbf{\Psi}^{s,t}\mathbf{y}\right) = \lim_{\tau\to 0}\frac{\mathbf{\Psi}^{s,t+\tau}\mathbf{y} - \mathbf{\Psi}^{s,t}\mathbf{y}}{\tau} \overset{\text{(iii)}}{=} \lim_{\tau\to 0}\frac{\mathbf{\Psi}^{t,t+\tau}(\mathbf{\Psi}^{s,t}\mathbf{y}) - \mathbf{\Psi}^{t,t}(\mathbf{\Psi}^{s,t}\mathbf{y})}{\tau} \overset{\text{(ii)}}{=} \mathbf{f}(t,\mathbf{\Psi}^{s,t}\mathbf{y})\ .$$

$t \mapsto \mathbf{\Psi}^{s,t}$ solves the same initial value problem for $\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y})$ as $t \mapsto \mathbf{\Phi}^{s,t}\mathbf{y}$. By the theorem of Picard-Lindelöf ($\rightarrow$ Thm. 1.3.4) there holds $\mathbf{\Psi} = \mathbf{\Phi}$.

Definition 2.1.8 (Consistency of a discrete evolution)**.**
The discrete evolution $\boldsymbol{\Psi}$ *is consistent with the ODE* $\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y})$*, if for all* $(t,\mathbf{y}) \in \Omega$

$$\boldsymbol{\Psi}^{t,t}\mathbf{y} = \mathbf{y} \quad \text{and} \quad \left.\frac{d}{ds}\boldsymbol{\Psi}^{t,t+s}\mathbf{y}\right|_{s=0} = \mathbf{f}(t,\mathbf{y}) \ .$$

Lemma 2.1.9 (Representation of consistent discrete evolutions)**.** $\rightarrow$ *[12, Lemma 4.4]*
Let $(t,\mathbf{y}) \in \Omega$ *and* $s \mapsto \boldsymbol{\Psi}^{t,t+s}\mathbf{y}$ *be* continuously differentiable *in the neighborhood of* 0*.*
$\boldsymbol{\Psi}$ *is consistent with* $\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y})$ *if and only if there exists an increment function* $h \mapsto \boldsymbol{\psi}(t,\mathbf{y},h)$ *that is continuous on this neighborhood of* 0 *and satisfies*

$$\boldsymbol{\Psi}^{t,t+h}\mathbf{y} = \mathbf{y} + h\boldsymbol{\psi}(t,\mathbf{y},h) \quad , \quad \psi(t,\mathbf{y},0) = \mathbf{f}(t,\mathbf{y}) \ . \tag{2.1.10}$$

Definition 2.1.11 (Consistency error of a discrete evolution)**.** $\rightarrow$ *[12, Def. 4.3]*

Consistency error: $\boldsymbol{\tau}(t,\mathbf{y},h) := \boldsymbol{\Phi}^{t,t+h}\mathbf{y} - \boldsymbol{\Psi}^{t,t+h}\mathbf{y}$ *(*h *sufficiently small);* .

Lemma 2.1.12 (Consistency and consistency error)**.**
Let $(t, \mathbf{y}) \in \Omega$, let $s \mapsto \mathbf{\Psi}^{t,t+s}\mathbf{y}$ be continuously differentiable in the neighborhood of 0. $\mathbf{\Psi}$ is consistent with $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$ ($\rightarrow$ Def. 2.1.8) if and only if the consistency error satisfies

$$\|\boldsymbol{\tau}(t, \mathbf{y}, h)\| = o(h) \quad \text{for } h \to 0 \quad \text{locally uniform in } (t, \mathbf{y}) \in \Omega \ .$$

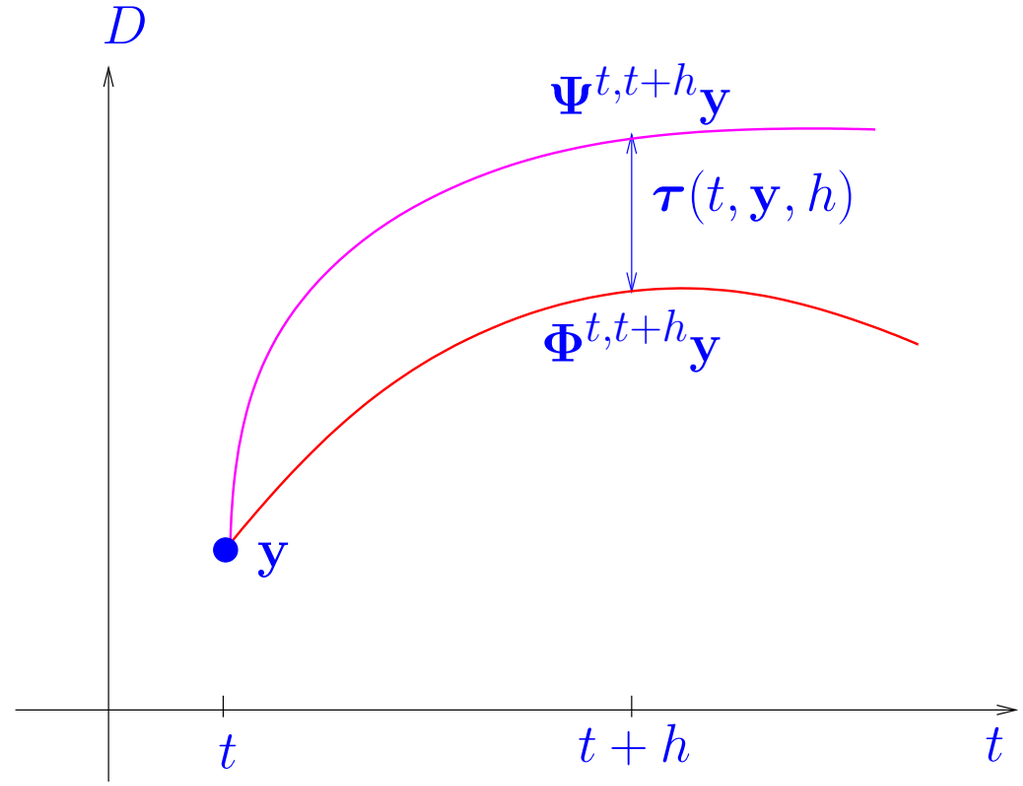
✎ Notation: "'Landau-o": $g(h) = o(h) \quad :\Leftrightarrow \quad \frac{g(h)}{h} \to 0$ for $h \to 0$

Interpretation:

consistency error = one step error

— $\hat{=}$ exact solution in $(y, \mathbf{y})$

— $\hat{=}$ approximate solution by discrete evolution



Definition 2.1.13 (Order of consistency of a discrete evolution). *[12, Def. 4.7]*

A discrete evolution has the order of consistency $p \in \mathbb{N}$, if the consistency error behaves locally uniform in Ω like

$$\|\boldsymbol{\tau}(t, \mathbf{y}, h)\| = O(h^{p+1}) \quad \text{for } h \to 0 \,. \tag{2.1.14}$$

P. Grohs

rev 63606, February 17, 2014

The statement that (2.1.14) holds *locally uniform* means

$$\forall (t,\mathbf{y}) \in \Omega\colon \quad \exists h_0, \delta, C > 0\colon \quad \tau(\widetilde{t},\widetilde{\mathbf{y}},h) \leq C h^{p+1} \quad \forall \widetilde{t},\widetilde{\mathbf{y}},h\colon \quad |\widetilde{t}-t| \leq \delta, \|\widetilde{\mathbf{y}}-\mathbf{y}\| \leq \delta \;, \quad 0 \leq h \leq h_0 \;.$$

Due to the equivalence of all norms on finite dimensional spaces $\mathbb{R}^d$, the choice of the norm in the definitions 2.1.11 and 2.1.13 does not make a difference.

Technique to determine the order of consistency: *Taylor expansion*

Example 2.1.15 (Order of consistency of simple single step methods).

Implicit midpoint rule (1.4.19): $\mathbf{y}_1 = \mathbf{y}_0 + h\mathbf{f}(\frac{1}{2}(t_0+t_1), \frac{1}{2}(\mathbf{y}_0+\mathbf{y}_1))$

Note: No explicit formula for $\boldsymbol{\Psi}$ **!** ("implicit" method)

For "lazy" people: computer algebra (MAPLE) **!**

```
D(y) := x -> f(y(x));
```

$$\mathsf{D}(y) := x \mapsto f(y(x))$$

```
y0 := y(0);
```

$$y0 := y(0)$$

```
solve(y0+h*f((y0+y1)/2)=y1,{y1});
```

$$\{y1 = RootOf(-y(0) - hf(1/2\,y(0) + 1/2\_Z) + \_Z)\}$$

```
assign(%);
Taylor(y1-y(h),h=0,4);
```

$$series\Big(\Big(-1/6\left(D^{(2)}\right)(f)(y(0))(f(y(0)))^2 - 1/6\,(\mathsf{D}(f)(y(0)))^2 f(y(0)) + 1/8\,f(y(0))\Big(\Big(D^{(2)}\Big)(f)(y(0))\,f(y(0)) + 2\,(\mathsf{D}(f)(y(0)))^2\Big)\Big)h^3 + O\left(h^4\right), h, 4\Big)$$

Implicit midpoint rule has order of consistency 2 **!**

◇

We always determine the *formal* order of consistency of a SSM under the assumption that the exact solution is sufficiently $(*)$ smooth !

$(*)$: "sufficiently" $\hat{=}$ as smooth as needed for the Taylor expansion

If the exact solution is not sufficiently smooth ☞ limited meaning of order of consistency for the actual behavior of the method.

In this course:

(Often) implicit assumption of "sufficient smoothness" !

2.1.3 Convergence

Consider the initial value problem

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}) \quad , \quad \mathbf{y}(t_0) = \mathbf{y}_0 \quad \text{for a} \quad (t_0, \mathbf{y}_0) \in \Omega := I \times D \; . \qquad 1.1.13$$

$f : \Omega \mapsto \mathbb{R}^d$ 1.3.2

1.3.4 $t \mapsto \mathbf{y}(t)$

Consider single step method ($\rightarrow$ Def. 2.1.2) with discrete evolution $\mathbf{\Psi} : \Omega_h \subset I \times I \times D \mapsto \mathbb{R}^d$

$$\mathbf{\Psi}^{t,t+h}\mathbf{y} := \mathbf{y} + h\boldsymbol{\psi}(t,\mathbf{y},h) \; . \tag{2.1.17}$$

Increment function

Assumption: local estimate for consistency error($\rightarrow$ Def. 2.1.11): for a $p \in \mathbb{N}$

$$\forall(\overline{t},\overline{\mathbf{y}}) \in \Omega\text{:} \quad \exists C_c > 0, \delta > 0\text{:} \quad \left\| \mathbf{\Phi}^{t,t+h}\mathbf{y} - \mathbf{\Psi}^{t,t+h}\mathbf{y} \right\| \leq C_c h^{p+1} \quad \forall |h| \text{ sufficiently small ,}$$
$$\forall t\,\mathbf{y}\text{:} \; |t-\overline{t}| < \delta, \; \|\mathbf{y}-\overline{\mathbf{y}}\| < \delta \; . \tag{2.1.18}$$

Note: Order of consistency p for discrete evolution $\mathbf{\Psi}$ w.r.t. $\dot{\mathbf{y}} = f(t,\mathbf{y})$ $\Rightarrow$ (2.1.18)

$\mathbf{y}_{\mathcal{G}} = (\mathbf{y}_k)_{k=0}^N$: grid function generated by SSM $\boldsymbol{\Psi}$ on time grid ($T > t_0 \hat{=}$ final time))

$$\mathcal{G} := \{t_0 < t_1 < \cdots < t_N = T\} \subset J(t_0, \mathbf{y}_0) \ ,$$

cmp. Def. 2.1.2.

Theorem 2.1.19 (Convergence theorem for single step method). *[12, Thm. 4.10]*
Let the assumption (2.1.18) *and the representation* (2.1.17) *hold true. If the increment function* $\boldsymbol{\psi}$ *is locally Lipschitz continuous (→ Def. 1.3.2) w.r.t. the state variable* $\mathbf{y}$*, then*

(i) the discrete evolution $\boldsymbol{\Psi}$ *provides for all time grids* $\mathcal{G}$ *with sufficiently small* $h_{\mathcal{G}}$ *a grid function* $\mathbf{y}_{\mathcal{G}}$ *for initial value* $\mathbf{y}_0$*,*

(ii) this family $\{\mathbf{y}_{\mathcal{G}}\}_{\mathcal{G}}$ *of grid functions of the order* p *converges to* $t \mapsto \mathbf{y}(t)$*, compare Def. 2.1.7*

Auxiliary tool for the proof:

Lemma 2.1.20 (Discrete Gronwall lemma, compare Lemma 1.3.29)**.**
If the series $(\xi_k)_{k\in\mathbb{N}_0}$, $\xi_k \geq 0$, *satisfies the difference inequality*

$$\xi_{k+1} \leq Ch_k^{p+1} + (1+Lh_k)\xi_k \;, \quad k \in \mathbb{N}_0 \;, \quad L, C, h_k \geq 0 \;, \tag{2.1.21}$$

then there holds

$$\xi_N \leq C\big(\max_{k=0,\dots,N-1} h_k^p\big)\frac{1}{L}\left(\exp\big(L\sum_{k=0}^{N-1} h_k\big) - 1\right) + \exp\big(L\sum_{k=0}^{N-1} h_k\big)\cdot\xi_0 \;, \quad N \in \mathbb{N}_0 \;.$$

Proof. (by induction w.r.t. N)

Using the convention that empty sums vanish, the claim is true for $N = 0$

Inductive step:

$$\begin{aligned}
\xi_{N+1} &\overset{(2.1.21)}{\leq} Ch_N^{p+1} + (1+Lh_N)\xi_N \\
&\overset{*}{\leq} Ch_N^{p+1} + (1+Lh_N)\left(C\big(\max_{k=0}^{N-1} h_k^p\big)\frac{1}{L}\left(\exp\big(L\sum_{k=0}^{N-1} h_k\big) - 1\right) + \exp\big(L\sum_{k=0}^{N-1} h_k\big)\xi_0\right) \\
&\overset{\blacksquare}{\leq} C\big(\max_{k=0}^{N} h_k^p\big)\left(h_N + \frac{1}{L}\Big(\exp\big(L\sum_{k=0}^{N} h_k\big) - 1 - Lh_N\Big)\right) + \exp\big(L\sum_{k=0}^{N} h_k\big)\xi_0
\end{aligned}$$

That's the claim of the lemma for $N+1$.

$*$: uses the induction assumption, i.e., the claim of the lemma for ξ_N.

■ uses the elementary estimate $1+x \leq \exp(x)$, $x \in \mathbb{R}$ (convexity of exponential function).

Proof of Thm. 2.1.19; Generalization of the proof of the algebraic convergence of the explicit Euler method from Section 1.4.1. The study of this section is recommended for preparation.

① Compact neighborhood of the solution trajectory $t \mapsto \mathbf{y}(t)$ for initial value $\mathbf{y}_0$:

$$K_\delta := \{(t,\mathbf{y}) \in I \times \mathbb{R}^d\colon \quad t_0 \leq t \leq T,\ \|\mathbf{y}-\mathbf{y}(t)\| \leq \delta\}\ , \quad \delta > 0\ .$$

for sufficiently small $\delta > 0$: $\quad K_\delta \subset \Omega$

due to the local Lipschitz condition for $\boldsymbol{\psi}$ and the local consistency estimate (2.1.18)

② Assumption **A1**: $(\mathbf{y}_k)_{k=0}^N$ exists and $\mathbf{y}_k \in K_\delta \subset \Omega$ for a $\delta > 0$. This assumption will be confirmed a posteriori (by induction w.r.t. N) for sufficiently small $h_{\mathcal{G}}$.

③ Note $\mathbf{y}_{k+1} = \boldsymbol{\Psi}^{t_k,t_{k+1}}\mathbf{y}_k$ ➢ recursion for error $\mathbf{e}_k := \mathbf{y}(t_k) - \mathbf{y}_k$

$$\mathbf{e}_{k+1} = \underbrace{\left(\mathbf{y}(t_{k+1}) - \boldsymbol{\Psi}^{t_k,t_{k+1}}\mathbf{y}(t_k)\right)}_{\text{one step error}} + \underbrace{\left(\boldsymbol{\Psi}^{t_k,t_{k+1}}\mathbf{y}(t_k) - \boldsymbol{\Psi}^{t_k,t_{k+1}}\mathbf{y}_k\right)}_{\text{propagated error}} \tag{2.1.22}$$

$$\overset{(2.1.17)}{=} \boldsymbol{\tau}(t_k,\mathbf{y}(t_k),h_k) + \mathbf{e}_k + h_k(\boldsymbol{\psi}(t_k,\mathbf{y}(t_k),h_k) - \boldsymbol{\psi}(t_k,\mathbf{y}_k,h_k)) \;,$$

where the definition 2.1.11 of the consistency error $\boldsymbol{\tau}$ has been used.

④ Arguments based on compactness:

- consequence of the local consistency error estimate (2.1.18): for $|h|$ sufficiently small

$$\exists C > 0: \quad \left\|\boldsymbol{\Phi}^{t,t+h}\mathbf{y} - \boldsymbol{\Psi}^{t,t+h}\mathbf{y}\right\| \leq C_c h^{p+1} \quad \forall (t,\mathbf{y}),(t+h,\mathbf{y}) \in K_\delta \;. \tag{2.1.23}$$

- consequence of the local Lipschitz continuity ($\rightarrow$ Def. 1.3.2) of the increment function $\boldsymbol{\psi}$: for $|h|$ sufficiently small

$$\exists L > 0: \quad \|\boldsymbol{\psi}(t,\mathbf{z},h) - \boldsymbol{\psi}(t,\mathbf{w},h)\| \leq L\,\|\mathbf{z}-\mathbf{w}\| \quad \forall (t,\mathbf{z}),(t,\mathbf{w}) \in K_\delta \;. \tag{2.1.24}$$

⑤ (2.1.22) & $\triangle$-inequality $\Rightarrow$ recursion for error norm:

$$\begin{aligned}\|\mathbf{e}_{k+1}\| &\leq \|\mathbf{e}_k\| + \|\boldsymbol{\tau}(t_k,\mathbf{y}(t_k),h_k)\| + h_k\,\|\boldsymbol{\psi}(t_k,\mathbf{y}(t_k),h_k) - \boldsymbol{\psi}(t_k,\mathbf{y}_k,h_k)\| \\ &\overset{(2.1.24)}{\leq} \|\mathbf{e}_k\| + \|\boldsymbol{\tau}(t_k,\mathbf{y}(t_k),h_k)\| + h_k L\,\|\mathbf{y}(t_k) - \mathbf{y}_k\|\end{aligned}$$

$$\overset{(2.1.23)}{\leq} \quad Ch_k^{p+1} + (1 + Lh_k)\,\|\mathbf{e}_k\| \ .$$

Application of the discrete Gronwall lemma with $\xi_k := \|\mathbf{e}_k\|$, $\xi_0 = 0$:

$$\text{Lemma 2.1.20} \quad \Rightarrow \quad \|\mathbf{e}_k\| \leq Ch_{\mathcal{G}}^p \frac{\exp(L(T - t_0)) - 1}{L} \ .$$

⑥ The estimate shows that $\mathbf{y}_k - \mathbf{y}(t_k) \to 0$ for $h_{\mathcal{G}} \to 0$. Therefore, one can prove by induction

$$\forall \delta > 0\colon \quad \exists h^* = h^*(\delta) > 0\colon \quad h_{\mathcal{G}} < h^* \quad \Rightarrow \quad (t_k, \mathbf{y}_k) \in K_\delta \quad \forall k \ .$$

Therefore, assumption **A1** is justified. □

Note: The proof only uses the *consistency error along the solution trajectory* $\boldsymbol{\tau}(t, \mathbf{y}(t), h)$. Consequently, we can weaken the prerequisites for the order of consistency p (→ Def. 2.1.13) to

$$\|\boldsymbol{\tau}(t, \mathbf{y}(t), h)\| \leq C_c h^{p+1} \quad \forall t \in [t_0, T] \ , \quad \text{for } h \text{ sufficiently small.}$$

Rule to remember: (Only) for single step methods:

order of consistency p $\Longrightarrow$ order of convergence p

The constant in the asymptotic error estimate in Thm. 2.1.19 depends *exponentially* on $T - t_0$, as a consequence of the discrete Gronwall lemma. This makes the estimate of the theorem potentially useless for *integrations over a long time span*, cmp. Lemma 4.4.82.

2.1.4 The equivalence principle (Dahlquist, Lax)

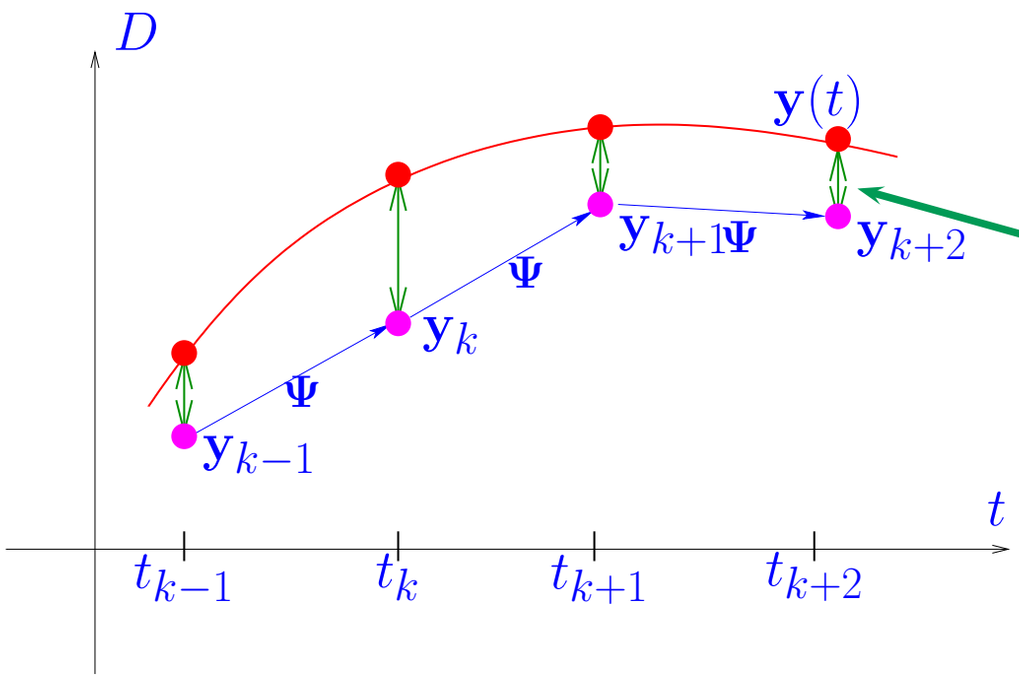
Goal: Abstraction of the proof of Thm. 2.1.19

Consider: equidistant time grid $\mathcal{G} = \{t_k\}_{k=0}^N$, $t_k := t_0 + hk$, $h := (T - t_0)/N$, $N \in \mathbb{N}$

discrete evolution $\boldsymbol{\Psi}$ $\longleftrightarrow$ (continuous) evolution $\boldsymbol{\Phi}$

grid function $\mathbf{y}_{\mathcal{G}} = (\mathbf{y}_k)_{k=0}^N$ solution $t \mapsto \mathbf{y}(t)$

(Assumption: $\mathcal{G} \subset J(t_0, \mathbf{y}_0)$, $\mathbf{y}_{\mathcal{G}}$ well-defined)



consistency error, cmp. Def. 2.1.11:

$$\boldsymbol{\tau}(t, \mathbf{y}, h) := (\boldsymbol{\Phi}^{t,t+h}\mathbf{y} - \boldsymbol{\Psi}^{t,t+h}\mathbf{y}) \ .$$

error function: $\mathbf{e}_k := \mathbf{y}(t_k) - \mathbf{y}_k$.

$\triangleleft$ — $\hat{=}$ $t \mapsto \mathbf{y}(t)$

• $\hat{=}$ $\mathbf{y}(t_k)$

• $\hat{=}$ $\mathbf{y}_k$

$\longrightarrow$ $\hat{=}$ $\boldsymbol{\Psi}^{t_k,t_{k+1}}$

P. Grohs

rev 63606, February 17, 2014

Error recursion

$$
\begin{aligned}
\mathbf{e}_k &= \mathbf{y}(t_k) - \mathbf{y}_k \\
&= \mathbf{y}(t_k) - \boldsymbol{\Psi}^{t_{k-1},t_k}(\mathbf{y}(t_{k-1}) - \mathbf{e}_{k-1}) \\
&= \underbrace{\boldsymbol{\Psi}^{t_{k-1},t_k}(\mathbf{y}(t_{k-1})) - \boldsymbol{\Psi}^{t_{k-1},t_k}(\mathbf{y}(t_{k-1}) - \mathbf{e}_{k-1})}_{\text{propagated error}} \\
&\quad + \underbrace{\mathbf{y}(t_k) - \boldsymbol{\Psi}^{t_{k-1},t_k}(\mathbf{y}(t_{k-1}))}_{\text{one step error = consistency error}}
\end{aligned}
$$

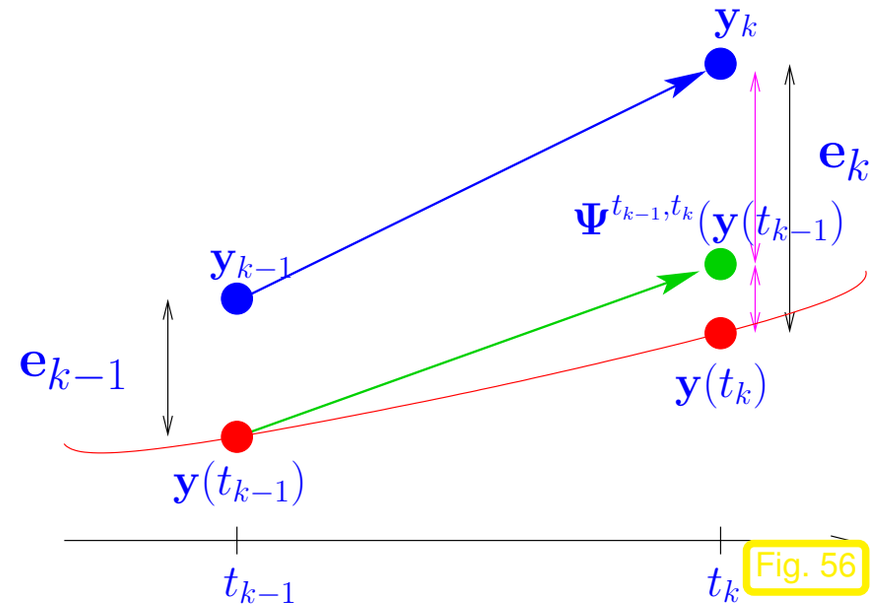


Fig. 56

Alternative error recursion

$$
\begin{aligned}
\mathbf{e}_k &= \mathbf{y}(t_k) - \mathbf{y}_k = \boldsymbol{\Phi}^{t_{k-1},t_k}\mathbf{y}(t_{k-1}) - \boldsymbol{\Psi}^{t_{k-1},t_k}(\mathbf{y}_{k-1}) \\
&= \underbrace{\boldsymbol{\Phi}^{t_{k-1},t_k}(\mathbf{y}_{k-1} + \mathbf{e}_{k-1}) - \boldsymbol{\Phi}^{t_{k-1},t_k}(\mathbf{y}_{k-1})}_{\text{propagated error}} + \underbrace{\boldsymbol{\Phi}^{t_{k-1},t_k}(\mathbf{y}_{k-1}) - \boldsymbol{\Psi}^{t_{k-1},t_k}(\mathbf{y}_{k-1})}_{\text{one step error = consistency error}}
\end{aligned}
$$

1. Error recursion: estimate of the consistency error in the neighborhood of the exact solution is sufficient

2. Error recursion: consistency error has to be estimated in the neighborhood of the solution of the SSM

Concept: Stability of a discrete evolution $\longleftrightarrow$ control of error propagation

Definition 2.1.25 (Nonlinear stability)**.**
A discrete evolution $\mathbf{\Psi}$ is (nonlinearly) stable

$$:\Leftrightarrow \quad \exists c > 0: \quad \left\| \mathbf{\Psi}^{t,t+h}\mathbf{y} - \mathbf{\Psi}^{t,t+h}\mathbf{z} \right\| \leq (1+ch)\left\|\mathbf{y}-\mathbf{z}\right\|$$

locally uniform in $(t,\mathbf{y})$ for sufficiently small $\|\mathbf{y}-\mathbf{z}\|$, $h>0$.

for SSM (2.1.17): local Lipschitz continuity of $\boldsymbol{\psi}$ $\succ$ nonlinear stability

Theorem 2.1.26 (consistency **&** (nonlinear) stability $\Rightarrow$ convergence)**.**
If $\mathbf{\Psi}$ consistent with $\mathbf{\Phi}$ (of order p) and (nonlinearly) stable, then the single step method converges globally (with order p).

2.1.5 Reversibility

We have seen: An approximate discrete evolution $\mathbf{\Psi}$ ($\rightarrow$ Sect. 2.1.1) can *in general* **not** satisfy $\mathbf{\Psi}^{r,s}\mathbf{\Psi}^{t,r} = \mathbf{\Psi}^{t,s}$ (otherwise the numerical method would return the exact solution).

However: for $s = t$, this request is feasible **!**

Definition 2.1.27 (Reversible discrete evolution). $\rightarrow$ *[12, Def. 4.40]*
A discrete evolution $\mathbf{\Psi} : \widetilde{\Omega}_h \subset I \times I \times D \mapsto \mathbb{R}^d$ *(and the respective single step method) is called reversible, if*

$$\mathbf{\Psi}^{t,s}\mathbf{\Psi}^{s,t}\mathbf{y} = \mathbf{y} \quad \forall (t,\mathbf{y}) \in \Omega \,, \quad \forall\, |t-s| \text{ sufficiently small} \,.$$

[Beitrag von J. Mitrovic, FS 2011]

Reversibility transfers an aspect of group properties from evolutions to the discrete world.

Example 2.1.28 (Simple reversible single step methods).

- implicit midpoint rule (1.4.19)

$$\mathbf{\Psi}^{t,t+h}\mathbf{y} = \mathbf{y} + h\mathbf{f}(t+\tfrac{1}{2}h, \tfrac{1}{2}(\mathbf{y}+\mathbf{\Psi}^{t,t+h}\mathbf{y}))$$
$$\Downarrow$$
$$\mathbf{y} = \mathbf{\Psi}^{t,t+h}\mathbf{y} - h\mathbf{f}(t+h-\tfrac{1}{2}h, \tfrac{1}{2}(\mathbf{y}+\mathbf{\Psi}^{t,t+h}\mathbf{y}) \;.$$

Assuming that we can uniquely solve the definition equation (1.4.19) for $\mathbf{y}_{k+1}$, there holds

$$\Rightarrow \qquad \mathbf{y} = \mathbf{\Psi}^{t+h,t}\mathbf{\Psi}^{t,t+h}\mathbf{y} \;.$$

- Störmer-Verlet method ($\rightarrow$ Sect. 1.4.4) in single step formulation from Remark 1.4.33

$$\begin{aligned} \mathbf{v}_{k+\frac{1}{2}} &= \mathbf{v}_k + \tfrac{h}{2}\mathbf{f}(\mathbf{y}_k) \;, \\ \mathbf{y}_{k+1} &= \mathbf{y}_k + h\mathbf{v}_{k+\frac{1}{2}} \;, \\ \mathbf{v}_{k+1} &= \mathbf{v}_{k+\frac{1}{2}} + \tfrac{h}{2}\mathbf{f}(\mathbf{y}_{k+1}) \;. \end{aligned} \qquad \Rightarrow \qquad \begin{aligned} \mathbf{v}_{k+\frac{1}{2}} &= \mathbf{v}_{k+1} - \tfrac{h}{2}\mathbf{f}(\mathbf{y}_{k+1}) \;, \\ \mathbf{y}_k &= \mathbf{y}_{k+1} - h\mathbf{v}_{k+\frac{1}{2}} \;, \\ \mathbf{v}_k &= \mathbf{v}_{k+\frac{1}{2}} - \tfrac{h}{2}\mathbf{f}(\mathbf{y}_k) \;. \end{aligned}$$

One can detect reversibility using the definition of the method: if switching $\mathbf{y}_k \leftrightarrow \mathbf{y}_{k+1}$ and $h \leftrightarrow -h$ results in the same equation, then the original method is reversible.

Theorem 2.1.29 (Order of consistency of reversible SSMs). $\rightarrow$ *[12, Satz 4.42]*

The maximal order of consistency ($\rightarrow$ Def. 2.1.13) of a reversible single step method ($\rightarrow$ Def. 2.1.27) is even.

The proof uses the following auxiliary tool ([12, Lemma 4.38]):

Lemma 2.1.30 (Perturbation lemma for discrete evolutions). *Let $\mathbf{f}$ be twice continuously differentiable and let $\mathbf{\Psi}$ be a discrete evolution that is consistent for the ODE $\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y})$ ($\rightarrow$ Def. 2.1.8) and continuously differentiable in h and $\mathbf{y}$. Then, there holds for $(t,\mathbf{y}) \in \Omega$ and sufficiently small $\mathbf{z} \in \mathbb{R}^d$*

$$\mathbf{\Psi}^{t,t+h}(\mathbf{y}+\mathbf{z}) = \mathbf{\Psi}^{t,t+h}\mathbf{y} + \mathbf{z} + h\frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t,\mathbf{y})\mathbf{z} + \mathbf{r}(h,\mathbf{z}) \;, \quad \|\mathbf{r}(h,\mathbf{z})\| \leq C(h^2\,\|\mathbf{z}\| + h\,\|\mathbf{z}\|^2) \;,$$

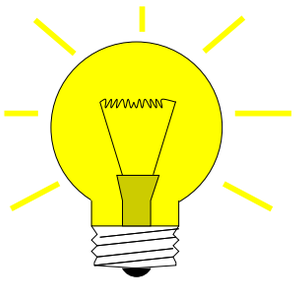
with $C > 0$ independent of h and $\mathbf{z}$.

☞ Thm. 2.1.29 explains the $O(h^2)$-convergence of the implicit midpoint rule that we observed in Ex. 1.4.21.

2.2 Collocation method[12, Sect. 6.3], [21, Sect. II.1.2]

2.2.1 Construction

First: Focus on first ($\leftrightarrow$ general) step
(this is sufficient for single step methods $\rightarrow$ Def. 2.1.2)



Idea: ❶ Approximate $\mathbf{y}(t)$, $t \in [t_0, t_1]$, in $s+1$ dimensional trial space V of functions $[t_0, t_1] \mapsto \mathbb{R}^d$ $\triangleright$ $\mathbf{y}_h$.

❷ Determination of $\mathbf{y}_h \in V$ by collocation conditions

$$\boxed{\mathbf{y}_h(t_0) = \mathbf{y}_0 \quad , \quad \dot{\mathbf{y}}_h(\tau_j) = \mathbf{f}(\tau_j, \mathbf{y}_h(\tau_j)) \; , \quad j = 1, \dots, s} \; , \qquad (2.2.1)$$

for collocation points $t_0 \leq \tau_1 < \dots < \tau_s \leq t_1$.

"Standard option": Polynomial trial space $V = \mathcal{P}_s$

✎ Notation: $\mathcal{P}_s \hat{=}$ space of univariant polynomial of degree $\leq s$, $s \in \mathbb{N}_0$

Known: $\dim \mathcal{P}_s = s+1$

➢ A polynomial $p \in \mathcal{P}_s$ is uniquely determined by $s+1$ interpolation conditions for values/derivatives.

➢ collocation conditions (2.2.1) imply polynomial degree s (in the sense of uniqueness/existence of $\mathbf{y}_h$)

Deduction: Formula for $\mathbf{y}_h(t_1)$ $(h := t_1 - t_0, \tau_j := t_0 + c_j h, 0 \le c_1 < c_2 < \ldots < c_s \le 1)$

Auxiliary tool: $\{L_j\}_{j=1}^s \subset \mathcal{P}_{s-1} \hat{=}$ Lagrange polynomials for points c_i, $i = 1, \ldots, s$, in $[0,1]$:

$$L_i(\tau) = \prod_{j=1, j\neq i}^{s} \frac{\tau - c_j}{c_i - c_j}\,, \quad i = 1, \ldots, s \quad \Rightarrow \quad L_j(c_i) = \delta_{ij}\,, \quad i, j = 1, \ldots, s\,. \tag{2.2.2}$$

(2.2.1) $\Rightarrow$

$$\dot{\mathbf{y}}_h(t_0 + \tau h) = \sum_{j=1}^{s} \mathbf{k}_j L_j(\tau) \quad , \quad \mathbf{k}_j := \mathbf{f}(t_0 + c_j h, \mathbf{y}_h(t_0 + c_j h))\,.$$

$$\Rightarrow \quad \mathbf{y}_h(t_0 + \tau h) = \mathbf{y}_0 + h \sum_{j=1}^{s} \mathbf{k}_j \int_0^\tau L_j(\zeta)\,\mathrm{d}\zeta\,.$$

Defining equation for the single step collocation method
(for collocation points $0 \le c_1 < c_2 < \cdots < c_s \le 1$):

$$\mathbf{y}_h(t_1) = \mathbf{y}_0 + h\sum_{i=1}^{s} b_i \mathbf{k}_i \;, \qquad \text{with} \qquad a_{ij} = \int_0^{c_i} L_j(\tau)\,\mathrm{d}\tau \;,$$

$$\mathbf{k}_i = \mathbf{f}(t_0 + c_i h, \mathbf{y}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j) \;. \qquad b_i = \int_0^{1} L_i(\tau)\,\mathrm{d}\tau \;. \tag{2.2.3}$$

Discrete evolution $\boldsymbol{\Psi}^{t_0,t_1} : \Omega \mapsto \Omega$, $\boldsymbol{\Psi}^{t_0,t_1}\mathbf{y}_0 := \mathbf{y}_1 := \mathbf{y}_h(t_1)$

➤ (2.2.3) $\hat{=}$ (Nonlinear) system of equations for increments $\mathbf{k}_i$ $(\approx \dot{\mathbf{y}}(t_0 + c_i h))$

▷ (Generic) collocation method = *implicit* single step method (→ Def. 2.1.2)

Single step collocation method in the form of Lemma 2.1.9:

$$\boldsymbol{\Psi}^{t_0,t_0+h}\mathbf{y}_0 = \mathbf{y}_0 + h\boldsymbol{\psi}(t_0,\mathbf{y}_0,h) \quad \text{with increment function} \quad \boldsymbol{\psi}(t_0,\mathbf{y}_0,h) = \sum_{i=1}^{s} b_i \mathbf{k}_i \;. \tag{2.2.4}$$

Remark 2.2.5 (Rephrasing of the increment equations (2.2.3)).

Equivalent form of the increment equations (2.2.3):

Replace increments $\mathbf{k}_i$ by $\boxed{\mathbf{g}_i := \mathbf{y}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j}$, $i = 1,\dots,s \quad \Leftrightarrow \quad \mathbf{k}_i = \mathbf{f}(t_0 + c_i h, \mathbf{g}_i)$.

$$(2.2.3) \quad \Leftrightarrow \quad \begin{aligned} \mathbf{g}_i &= \mathbf{y}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{f}(t_0 + c_j h, \mathbf{g}_j) \\ \mathbf{y}_1 &= \mathbf{y}_0 + h\sum_{i=1}^{s} b_i\mathbf{f}(t_0 + c_i h, \mathbf{g}_i)\ . \end{aligned} \tag{2.2.6}$$

△

Lemma 2.2.7 (Solvability of increment equations).
Let $\mathbf{f}$ be locally Lipschitz continuous ($\rightarrow$ Def. 1.3.2) on the augmented state space Ω. Then, for every $(t_0, \mathbf{y}_0) \in \Omega$, there exists a $h_0 > 0$ such that (2.2.3) *is uniquely solvable for the increments for every $h < h_0$, and such that these are continuous functions in h.*
For $f \in C^m(\Omega, \mathbb{R}^d)$, $m \in \mathbb{N}$, these increments are m-times continuously differentiable functions of $\mathbf{y}_0, t_0, h$.

"Proof" (of Lemma 2.2.7, *using stricter smoothness prerequisites*, here only given for the autonomous case $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$)

Assumption: $\mathbf{f}$ is continuously differentiable on Ω

$$\mathbf{k}_i = \mathbf{f}(\mathbf{y}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j) \quad \Leftrightarrow \quad G(h, \mathfrak{k}) = 0\,, \quad G(h, \mathfrak{k}) := \mathfrak{k} - \begin{pmatrix} \mathbf{f}(\mathbf{y}_0 + h\sum_{j=1}^{s} a_{1j}\mathbf{k}_j) \\ \vdots \\ \mathbf{f}(\mathbf{y}_0 + h\sum_{j=1}^{s} a_{sj}\mathbf{k}_j) \end{pmatrix}\,,$$

with $\mathfrak{k} = (\mathbf{k}_1, \ldots, \mathbf{k}_s)^T \in \mathbb{R}^{s\cdot d}$.
Idea: Application of implicit function theorem on $G : \mathbb{R} \times D \mapsto D$

Theorem 2.2.8 (Implicit function theorem). $\rightarrow$ *analysis course*

Let $I \subset \mathbb{R}^q$, $U \subset \mathbb{R}^n$ *be open and* $G = G(\xi, y) : I \times U \mapsto \mathbb{R}^n$ *be continuously differentiable.*
For a $(\xi_0, y_0) \in I \times U$ *let* $G(\xi, y) = 0$.
If the Jacobian matrix $\frac{\partial G}{\partial y}(\xi_0, y_0)$ *is invertible, then there exists a neighborhood* $V \subset I$ *of* ξ_0 *and a unique continuously differentiable function* $\xi \mapsto z(\xi)$ *such that*

$$G(\xi, z(\xi)) = 0 \quad \forall \xi \in V \ .$$

- $\mathfrak{k}_0 := (\mathbf{f}(\mathbf{y}_0), \ldots, \mathbf{f}(\mathbf{y}_0))^T$ satisfies $\quad G(0, \mathfrak{k}_0) = 0$
- Derivative ($\hat{=}$ Jacobian matrix) of G in $(0, \mathfrak{k}_0)$ (chain rule)

$$D_{\mathfrak{k}} G(0, \mathfrak{k}_0) = \mathbf{I}$$

is the unity matrix and therefore obviously differentiable. $\square$

An alternative, technically more complicated proof requires only local Lipschitz continuity of $\mathbf{f}$ and produces additional bounds on the step size for the existence of a solution of the increment equations:

Auxiliary tool for alternative proof ($\rightarrow$ analysis course):

Theorem 2.2.9 (Banach fixed-point theorem, parameter dependent version)**.**
Let $V \subset \mathbb{R}^d$ closed, $U \subset \mathbb{R}^n$ open, let $F : U \times V \mapsto V$ be (totally) m-times continuously differentiable, $m \in \mathbb{N}_0$, and let there hold the uniform contraction property

$$\exists 0 \leq q < 1: \quad \|F(\mathbf{u},\mathbf{z}) - F(\mathbf{u},\mathbf{w})\| \leq q\,\|\mathbf{z}-\mathbf{w}\| \quad \forall \mathbf{z},\mathbf{w} \in V, \quad \forall \mathbf{u} \in U \ .$$

Then there exists a m-times continuously differentiable function $G : U \mapsto V$ such that

$$F(\mathbf{u}, G(\mathbf{u})) = G(\mathbf{u}) \quad \forall \mathbf{u} \in U \ .$$

✎ Standard notation for coefficient matrix $\mathfrak{A} := \left(a_{ij}\right)_{i,j=1}^{s} \in \mathbb{R}^{s,s}$

✎ Matrix norm induced by maximum norm: $\|\mathfrak{A}\|_\infty := \max\limits_{i=1,\dots,s} \sum\limits_{j=1}^{s} |a_{ij}|$

Proof. (of Lemma 2.2.7 for autonomous case $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$)

Preparation: Like in the proof of Thm. 2.1.19 we consider $\mathbf{f}$ on the compact neighborhood K_δ of the solution curve $t \mapsto \mathbf{y}(t)$ on the augmented state space Ω. Therefore (for now) WLOG we *assume*:

$\mathbf{f}$ globally Lipschitz continuous, cmp. Def. 1.3.2:

$$\exists L > 0: \quad \|\mathbf{f}(\mathbf{z}) - \mathbf{f}(\mathbf{w})\| \leq L \|\mathbf{z} - \mathbf{w}\| \quad \forall \mathbf{z}, \mathbf{w} \in D \; . \tag{2.2.10}$$

We also assume that an r-neighborhood of $\mathbf{y}_0$ is contained in D:

$$\exists r > 0: \quad \|\mathbf{z} - \mathbf{y}_0\| \leq r \quad \Rightarrow \quad \mathbf{z} \in D \; .$$

Idea: Application of Banach fixed-point theorem 2.2.9 on the (equivalent) increment equations (2.2.6) for the $\mathbf{g}_i$: with $\mathfrak{g} := (\mathbf{g}_1, \dots, \mathbf{g}_s) \in \mathbb{R}^{s \cdot d}$

$$(2.2.6) \quad \Leftrightarrow \quad \mathfrak{g} = F(h, \mathfrak{g}) \; , \quad F(h, \mathfrak{g}) := \begin{pmatrix} \mathbf{y}_0 + h \sum_{j=1}^{s} a_{1j} \mathbf{f}(\mathbf{g}_j) \\ \vdots \\ \mathbf{y}_0 + h \sum_{j=1}^{s} a_{sj} \mathbf{f}(\mathbf{g}_j) \end{pmatrix} \; . \tag{2.2.11}$$

Norm used on $\mathbb{R}^{s\cdot d}$: $\|\mathfrak{g}\| := \max_{i=1,\dots,s} \|\mathbf{g}_i\|$.

To show: all $\mathbf{g}_i$ stay in the *closed* r-neighborhood of $\mathbf{y}_0$ for sufficiently small h: with $\mathfrak{y}_0 = (\mathbf{y}_0, \dots, \mathbf{y}_0)$

$$\|F(h,\mathfrak{g}) - \mathfrak{y}_0\| = \max_{i=1,\dots,s} |h| \left\| \sum_{j=1}^{s} a_{ij}\mathbf{f}(\mathbf{g}_j) \right\| \leq |h| \, \|\mathfrak{A}\|_\infty \max_{j=1,\dots,s} \|f(\mathbf{y}_0) + f(\mathbf{g}_j) - f(\mathbf{y}_0)\|$$

$$\overset{(2.2.10)}{\leq} |h| \, \|\mathfrak{A}\|_\infty \left(\|\mathbf{f}(\mathbf{y}_0)\| + L\,\|\mathfrak{g} - \mathfrak{y}_0\|\right) \ .$$

$$\Rightarrow \quad \left\{ |h| < \frac{r}{\|\mathfrak{A}\|_\infty \left(\|\mathbf{f}(\mathbf{y}_0)\| + Lr\right)} \quad \Rightarrow \quad \|F(h,\mathfrak{g}) - \mathfrak{y}_0\| \leq r \text{ , if } \|\mathfrak{g} - \mathfrak{y}_0\| \leq r \right\} \ .$$

To show: $\mathfrak{g} \mapsto F(h,\mathfrak{g})$ is a contraction that is uniform in h

$$\begin{aligned} \|F(h,\mathfrak{g}) - F(h,\mathfrak{p})\| &\leq |h| \cdot \max_{i=1,\dots,s} \left\| \sum_{j=1}^{s} a_{ij}(\mathbf{f}(\mathbf{g}_j) - \mathbf{f}(\mathbf{p}_j)) \right\| \\ &\leq |h| \cdot \|\mathfrak{A}\|_\infty \max_{j=1,\dots,s} \|\mathbf{f}(\mathbf{g}_j) - \mathbf{f}(\mathbf{p}_j)\| \\ &\leq |h| L \cdot \|\mathfrak{A}\|_\infty \|\mathfrak{g} - \mathfrak{p}\| \ , \end{aligned}$$

where, in the last step, we used the global Lipschitz condition for $\mathbf{f}$.

$$|h| < \frac{1}{L\,\|\mathfrak{A}\|_\infty} \quad \Rightarrow \quad \mathfrak{g} \mapsto F(h,\mathfrak{g}) \text{ is an } h\text{-uniform contraction.}$$

Choose
$$\boxed{h_0 = \frac{r}{\|\mathfrak{A}\|_\infty \left(\|\mathbf{f}(\mathbf{y}_0)\| + Lr\right)}}$$

Then, F fulfills with $U =]-h_0, h_0[$ and $V = \{\mathfrak{g}\colon \|\mathfrak{g} - \mathfrak{y}_0\| \leq r\}$ the prerequisites of the Banach fixed-point theorem 2.2.9.

$$\Rightarrow \quad \left\{\mathbf{f} \in C^m(D) \quad \Rightarrow \quad \exists \mathfrak{g} :]-h_0, h_0[\mapsto \mathbb{R}^{d\cdot s}\colon \quad F(h, \mathfrak{g}(h)) = \mathfrak{g}(h)\right\} .$$

Due to the equivalence (2.2.11), this concludes the proof. □

P. Grohs

rev 63606, February 17, 2014

Remark 2.2.12 (Bound on time step size from Lemma 2.2.7).

Based on the proof of Lemma 2.2.7 by means of the fixed-point argument, Thm 2.2.9, the solvability of the increment equations is only guaranteed if

$$|h| \leq \frac{1}{L\,\|\mathfrak{A}\|_\infty} ,$$

where $L > 0$ is a (local) Lipschitz constant ($\rightarrow$ Def. 1.3.2) for the source term $\mathbf{f}$.

This is a bound on the step size analogous to the bound on the step size for the explicit Euler method

near strongly attractive fixed-points, cmp. Ex. 1.4.9.

It still remains to verify the prerequisites of Thm. 2.1.19:

Lemma 2.2.13 (Lipschitz continuity of the increment function)**.**
Under the prerequisites of Lemma 2.2.7, there exists for every $(t_0, \mathbf{y}_0) \in \Omega$ *a* $h_0 > 0$ *such that* $\boldsymbol{\psi}$ *from* (2.2.4) *is locally Lipschitz continuous in the state space.*

Proof is based on the implicit function theorem, Thm. 2.2.8, assuming sufficient smoothness of $\mathbf{f}$:

Like in the first proof of Lemma 2.2.7, rephrase the increment equations as parameter dependent

root-finding problem:

$$\mathbf{k}_i = \mathbf{f}(\mathbf{y}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j) \quad \Leftrightarrow \quad G(h,\mathbf{y}_0,\mathfrak{k}) = 0\,, \quad G(h,\mathbf{y}_0,\mathfrak{k}) := \mathfrak{k} - \begin{pmatrix} \mathbf{f}(\mathbf{y}_0 + h\sum_{j=1}^{s} a_{1j}\mathbf{k}_j) \\ \vdots \\ \mathbf{f}(\mathbf{y}_0 + h\sum_{j=1}^{s} a_{sj}\mathbf{k}_j) \end{pmatrix}\,,$$

There holds

- G is continuously differentiable w.r.t. all arguments, if $\mathbf{f}$ is sufficiently smooth.
- $G(0,\mathbf{y}_0,\mathfrak{k}) = 0$ for $\mathfrak{k} = (\mathbf{f}(\mathbf{y}_0),\dots,\mathbf{f}(\mathbf{y}_0)) \in \mathbb{R}^{s\cdot d}$
- $D_{\mathfrak{k}}G(0,\mathbf{y}_0,\mathfrak{k}) = \mathbf{I}$ for arbitrary $\mathfrak{k} \in \mathbb{R}^{s\cdot d}$, $\mathbf{y}_0 \in D$.

Due to the implicit function theorem, there exists a locally continuously differentiable solution curve $\mathfrak{k} = \mathfrak{k}(\mathbf{y},h)$, defined in the neighborhood of $(0,\mathbf{y}_0)$. Therefore, the claim follows from an analogon to Lemma 1.3.3. □

Proof of Lemma 2.2.13 with fixed-point argument, without requirements on the smoothness of $\mathbf{f}$ (for autonomous ODE):

WLOG, $\mathbf{f}$ is assumed to be globally Lipschitz continuous, cmp. proof of Lemma 2.2.7:

$$\exists L > 0: \quad \|\mathbf{f}(\mathbf{z}) - \mathbf{f}(\mathbf{w})\| \leq L \|\mathbf{z} - \mathbf{w}\| \quad \forall \mathbf{z}, \mathbf{w} \in D \; . \tag{2.2.14}$$

With $\mathbf{g}_i$ from the equivalent increment equations (2.2.6):

$$\text{Increment function:} \quad \boldsymbol{\psi}(t, \mathbf{y}, h) = \mathbf{y} + h \sum_{j=1}^{s} b_i \mathbf{f}(\mathbf{g}_j) \quad , \quad \mathbf{g}_i = \mathbf{y} + h \sum_{j=1}^{s} a_{ij} \mathbf{f}(\mathbf{g}_j) \; . \tag{2.2.15}$$

Choose $\mathbf{y}, \mathbf{z} \in D$ and define (for sufficiently small h, compare Lemma 2.2.7) $\mathbf{g}_i^y, \mathbf{g}_i^z \in \mathbb{R}^d$ as solutions of

$$\left.\begin{aligned} \mathbf{g}_i^y &= \mathbf{y} + \sum_{j=1}^{s} a_{ij} \mathbf{f}(\mathbf{g}_j^y) \; , \\ \mathbf{g}_i^z &= \mathbf{z} + \sum_{j=1}^{s} a_{ij} \mathbf{f}(\mathbf{g}_j^z) \; . \end{aligned}\right. \quad , \quad i = 1, \dots, s \; .$$

With $\mathfrak{g}^y := (\mathbf{g}_1^y, \dots, \mathbf{g}_s^y)$, $\mathfrak{g}^z := (\mathbf{g}_1^z, \dots, \mathbf{g}_s^s)$:

$$\|\mathfrak{g}^y - \mathfrak{g}^z\| \leq \|\mathbf{y} - \mathbf{z}\| + h \max_{i=1,\dots,s} \sum_{j=1}^{s} |a_{ij}| \left(\left\| \mathbf{f}(\mathbf{g}_j^y) \right\| - \left\| \mathbf{f}(\mathbf{g}_j^z) \right\| \right)$$

$$\overset{(2.2.14)}{\leq} \quad \|\mathbf{y}-\mathbf{z}\| + hL \cdot \|\mathfrak{A}\|_\infty \|\mathfrak{g}^y - \mathfrak{g}^z\| \ .$$

$$h\,\|\mathfrak{A}\|_\infty L < 1 \quad \Rightarrow \quad \|\mathfrak{g}^y - \mathfrak{g}^z\| \leq \frac{1}{1-h\,\|\mathfrak{A}\|_\infty L}\,\|\mathbf{y}-\mathbf{z}\| \ .$$

cmp. the bound on the step size from Rem. 2.2.12

From this estimate and again with (2.2.14), it follows that (if $h\,\|\mathfrak{A}\|_\infty L < 1$)

$$\|\boldsymbol{\psi}(t,\mathbf{y},h) - \boldsymbol{\psi}(t,\mathbf{z},h)\| \leq h\sum_{i=1}^{s}|b_i|\,\left\|\mathbf{f}(\mathbf{g}_i^y) - \mathbf{f}(\mathbf{g}_i^z)\right\| \leq \frac{L}{1-Lh\,\|\mathfrak{A}\|_\infty}\sum_{i=1}^{s}|b_i|\cdot\|\mathbf{y}-\mathbf{z}\| \ . \tag{2.2.16}$$

As $\mathbf{y}$, $\mathbf{z}$ arbitrary, this implies the claim. □

P. Grohs

rev 63606, February 17, 2014

Example 2.2.17 (Convergence of simple single step collocation method).

- Scalar logistic differential equation (1.2.2), $\lambda = 10$, $y(0) = 0.01$, $T = 1$
- Single step collocation method (2.2.3) for $s = 1,\dots,4$, uniform step size h

Equidistant collocation points:

$$
\begin{aligned}
s = 1 : \ \mathbf{c} &= (\tfrac{1}{2}) \ , \\
s = 2 : \ \mathbf{c} &= (\tfrac{1}{3}, \tfrac{2}{3})^T \ , \\
s = 3 : \ \mathbf{c} &= (\tfrac{1}{4}, \tfrac{1}{2}, \tfrac{3}{4})^T ; , \\
s = 4 : \ \mathbf{c} &= (\tfrac{1}{5}, \tfrac{2}{5}, \tfrac{3}{5}, \tfrac{4}{5})^T \ .
\end{aligned}
$$

Numerical rates of convergence :
(computed by linear regression)

$$
\begin{aligned}
s = 1 \ &: \ p = 1.96 \\
s = 2 \ &: \ p = 2.03 \\
s = 3 \ &: \ p = 4.00 \\
s = 4 \ &: \ p = 4.04
\end{aligned}
$$

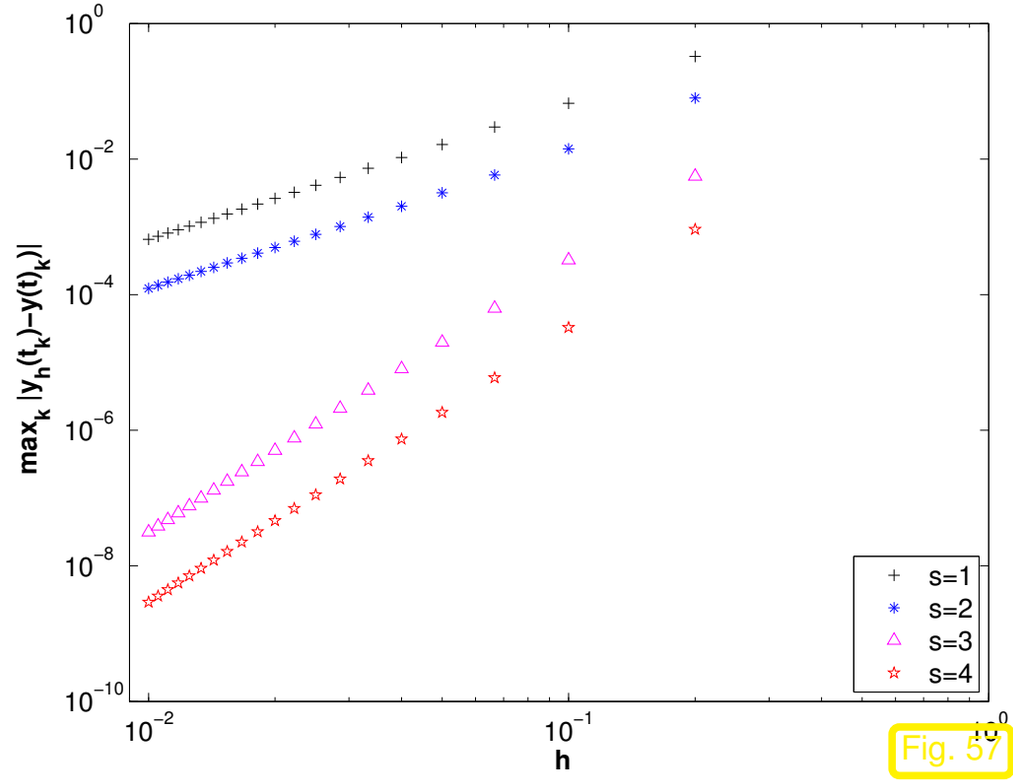


Fig. 57

P. Grohs

rev 63606, February 17, 2014

Translated equidistant collocation points:

$$
\begin{aligned}
s=1: \ & \mathbf{c}=(0)\ , \\
s=2: \ & \mathbf{c}=(0,\tfrac{1}{2})^T\ , \\
s=3: \ & \mathbf{c}=(0,\tfrac{1}{3},\tfrac{2}{3})^T;\ , \\
s=4: \ & \mathbf{c}=(0,\tfrac{1}{4},\tfrac{1}{2},\tfrac{3}{4})^T\ .
\end{aligned}
$$

Numerical rates of convergence :
(computed by linear regression)

$$
\begin{aligned}
s=1 \ &: \ p=0.95 \\
s=2 \ &: \ p=1.77 \\
s=3 \ &: \ p=2.95 \\
s=4 \ &: \ p=3.92
\end{aligned}
$$

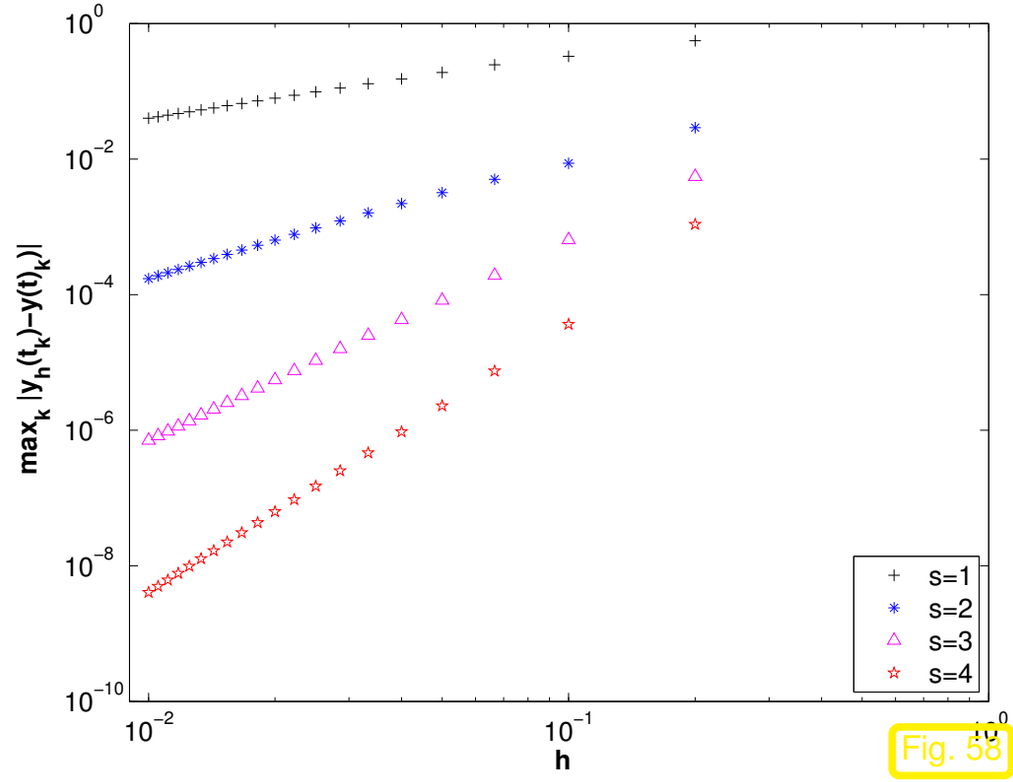


Fig. 58

Observation for *symmetrically* chosen, equidistant collocation points: Algebraic convergence of the order

$$
\begin{cases} p=s+1 & \text{, if } s \text{ odd ,} \\ p=s & \text{, if } s \text{ even.} \end{cases}
$$

Explanation $\rightarrow$ Sect. 2.2.3 & Thm. 2.1.29

$\diamond$

Remark 2.2.18 (Collocation method and numerical quadrature).

$\mathbf{f}(t,\mathbf{y}) = \mathbf{f}(t)$ & $\mathbf{y}_0 = 0$ ➤ Numerical quadrature ($\rightarrow$ course "Numerical methods")

$$\mathbf{y}(t_1) = \int_{t_0}^{t_1} \mathbf{f}(t)\,\mathrm{d}t \approx h \sum_{i=1}^{s} b_j \mathbf{f}(t_0 + c_j h) \quad = \text{Quadrature formula}$$

➞ $c_1, \ldots, c_s \leftrightarrow$ nodes of a quadrature formula (e.g., Gauss points on $[0,1]$)

$b_1, \ldots, b_s \leftrightarrow$ weights of a quadrature formula

△

P. Grohs

rev 63606, February 17, 2014

From the connection between collocation methods and numerical quadrature

➢ Choice of collocation points c_i as nodes of well-established quadrature formulae on $[0,1]$

The following examples show that this leads to useful methods:

- Case $s = 1$ & $c_1 = 1/2$ ($\leftrightarrow$ easiest Gauss-Legendre quadrature formula)

$$L_1 \equiv 1 \quad \Rightarrow \quad a_{11} = 1/2 \,, \quad b_1 = 1 \,.$$
$$\mathbf{k}_1 = \mathbf{f}(t_0 + 1/2h, \mathbf{y}_0 + 1/2h\mathbf{k}_1) \quad , \quad \mathbf{y}_h(t_1) = \mathbf{y}_0 + h\mathbf{k}_1 \,. \tag{2.2.19}$$

(2.2.19) = implicit midpoint rule (1.4.19)

- Case $s = 1$ & $c_1 = 0$ ($\leftrightarrow$ sinistral one-point-quadrature formula)

$$L_1 \equiv 1 \quad \Rightarrow \quad a_{11} = 0 \,, \quad b_1 = 1 \,.$$
$$\mathbf{k}_1 = \mathbf{f}(t_0, \mathbf{y}_0) \quad , \quad \mathbf{y}_h(t_1) = \mathbf{y}_0 + h\mathbf{k}_1 = \mathbf{y}_0 + h\mathbf{f}(t_0, \mathbf{y}_0) \,.$$

(2.2.1) = explicit Euler method (1.4.2) (not necessary to solve an equation !)

- Case $s = 1$ & $c_1 = 1$ ($\leftrightarrow$ dexter one-point-quadrature formula)

$$L_1 \equiv 1 \quad \Rightarrow \quad a_{11} = 1 \,, \quad b_1 = 1 \,.$$
$$\mathbf{k}_1 = \mathbf{f}(t_1, \mathbf{y}_0 + h\mathbf{k}_1) \quad , \quad \mathbf{y}_h(t_1) = \mathbf{y}_0 + h\mathbf{k}_1 = \mathbf{y}_0 + h\mathbf{f}(t_1, \mathbf{y}_h(t_1)) \,.$$

(2.2.1) = implicit Euler method

Reminder: "optimal quadrature method": Gaussian quadrature

($\rightarrow$ course "Numerical methods"[13, Sect. 9.3])

- ☞ The n nodes of the nth Gaussian quadrature formula on $[-1,1]$, $n \in \mathbb{N}$, are the roots of the Legendre polynomial of degree n ▷
- ☞ The nth Gaussian quadrature formula has order $2n$.
- ☞ The weights of the nth Gaussian quadrature formula are positive.

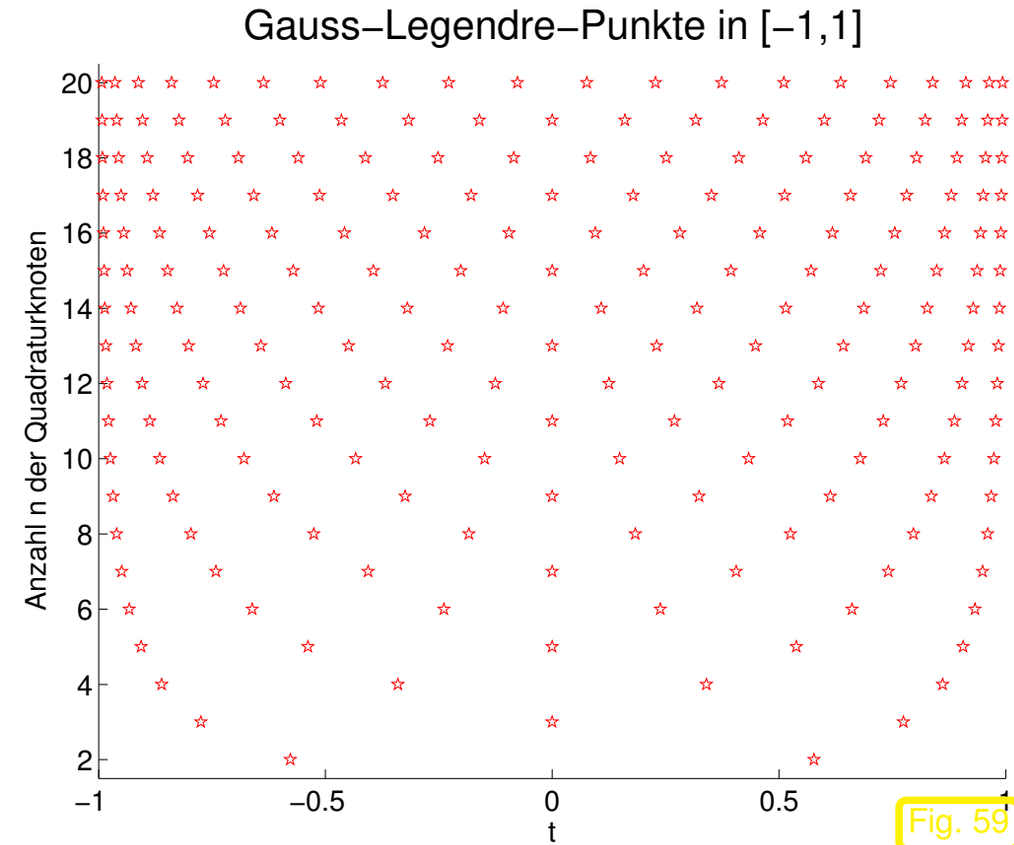


Fig. 59

P. Grohs

rev 63606, February 17, 2014

Single step Gauss collocation method

Remark 2.2.20 (Solution *function* from collocation method).

▷ Collocation methods produce (per constructionem) even *piecewise polynomial* approximate solutions $\mathbf{y}_h \in C^0([t_0,T])$

2.2.2 Abstract projection method

The goal of this section is to identify the collocation method as member of a bigger class of single step methods that allow for an elegant, abstract convergence theory.

Remark 2.2.21 (Collocation method as projection method).

$\mathsf{P}_s : C^0([t_0,t_1]) \mapsto \mathcal{P}_{s-1} \hat{=}$ Polynomial interpolation operator for nodes $\tau_1 \leq \cdots \leq \tau_s$ (cmp. Sect. 2.2.1, "collocation points")

Using this, we can rephrase the collocation conditions in a compact way:

$$\Rightarrow \quad \Big((2.2.1) \quad \Leftrightarrow \quad \dot{\mathbf{y}}_h = \mathsf{P}_s \mathbf{f}(\cdot, \mathbf{y}_h(\cdot)) \quad , \quad \mathbf{y}_h(t_0) = \mathbf{y}_0 \, . \Big)$$

Note: projection property $\mathsf{P}_s^2 = \mathsf{P}_s$

Known from linear algebra:

Definition 2.2.22 (Projection operator)**.**
Let X be a vector space. A linear *mapping $\mathsf{P} : X \mapsto X$ is a projection operator, if $\mathsf{P}^2 = \mathsf{P}$.*

Known from analysis:

Definition 2.2.23 (Continuous linear operator)**.**
Let X, Y be normed vector spaces. A linear operator $\mathsf{T} : X \mapsto Y$ is called continuous/bounded, if

$$\|\mathsf{T}\| := \sup_{x \in X \setminus \{0\}} \frac{\|\mathsf{T}x\|_Y}{\|x\|_X} < \infty \,.$$

$\|\mathsf{T}\|$ is called the norm of the continuous operator T.

Consider: ODE $\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y})$, $\mathbf{f} : I \times D \mapsto \mathbb{R}^d$ locally Lipschitz continuous, compare Sect. 1.1
Corresponding IVP $\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y})$, $\mathbf{y}(t_0) = \mathbf{y}_0$, on $[t_0, T] \in J(t_0, \mathbf{y}_0)$

► Generalization of collocation-SSM : single step projection method $\boldsymbol{\Psi}^{t,t+h}$ for ODE $\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y})$

discrete evolution of SSM

$\boldsymbol{\Psi}^{t,t+h}\mathbf{y}_0$ defined by

- finite dimensional trial space

$$V \subset C^1([t,t+h],\mathbb{R}^d) \quad ➥ \quad W := \{\frac{d}{dt}\mathbf{v} : \mathbf{v} \in V\}$$

- continuous projection operator $\mathsf{P} : C^0([t,t+h],\mathbb{R}^d) \mapsto W$

$$\boldsymbol{\Psi}^{t,t+h}\mathbf{y}_0 := \mathbf{y}_h(t+h) \quad \text{with} \quad \mathbf{y}_h \in V \quad \wedge \quad \begin{array}{l} \dot{\mathbf{y}}_h = \mathsf{P}(\mathbf{f}(\cdot,\mathbf{y}_h(\cdot))) \;, \\ \mathbf{y}_h(t) = \mathbf{y}_0 \in D \;. \end{array} \tag{2.2.24}$$

interpreted as function $\in C^0([t,t+h],\mathbb{R}^d)$

P. Grohs

rev 63606, February 17, 2014

Remark 2.2.25 (Fixed-point form of single step projection method).

Generalization: (2.2.24) $\longrightarrow$ fixed-point equation,

$$\text{(2.2.24)} \quad \Rightarrow \quad \mathbf{y}_h(\tau) = \mathbf{y}_0 + \int_t^\tau \mathsf{P}(\mathbf{f}(\cdot, \mathbf{y}_h(\cdot)))(\xi)\,\mathrm{d}\xi\,, \quad t \le \tau \le t+h\,. \tag{2.2.26}$$

! less strict smoothness assumptions for V: (2.2.26) useful for $\mathbf{y}_h \in (C^0([t,t+h]))^d$.

△

Lemma 2.2.27 (Requirement for the discrete evolution for single step projection methods being well-defined)**.**
Let $\mathbf{f}$ be locally Lipschitz continuous ($\rightarrow$ Def. 1.3.2). Then, $\mathbf{\Psi}^{t,t+h}\mathbf{y}_0$ is well-defined for h sufficiently small.

✎ Notation: Maximum norm $\|\mathbf{y}(\cdot)\|_{\infty,I} := \max_{\tau\in I} \|\mathbf{y}(\tau)\|$,
in particular in the following: $\|\mathbf{y}(\cdot)\|_\infty := \max_{t\le\tau\le t+h} \|\mathbf{y}(\tau)\|$

Proof. (Analogous to the proof of lemma 2.2.7, contraction argument)

We need to show the unique solvability of the equation (2.2.24) for the function $\mathbf{y}_h$.

Technique: Banach fixed-point theorem 2.2.9 applied to the fixed-point equation (2.2.26), cmp. proof of Thm. 1.3.4.

$$\mathbf{y}_h(\tau) = F(\mathbf{y}_h)\ , \quad F(\mathbf{y}_h)(\tau) := \mathbf{y}_0 + \int_t^\tau \mathsf{P}(\mathbf{f}(\cdot,\mathbf{y}_h(\cdot)))(\xi)\,\mathrm{d}\xi\ , \quad t \le \tau \le t+h\ . \tag{2.2.28}$$

Note: property of map: $F : (C^0([t,t+h]))^d \mapsto (C^0([t,t+h]))^d$

Reminder of analysis: ➢ $(C^0([t,t+h]), \|\cdot\|_\infty)$ is Banach space !

P. Grohs

rev 63606, February 17, 2014

Local Lipschitz condition & compactness argument, cmp. proof of Thm. 2.1.19

$$\exists L > 0\colon \quad \|\mathbf{f}(\tau,\mathbf{z}) - \mathbf{f}(\tau,\mathbf{w})\| \le L\,\|\mathbf{z}-\mathbf{w}\| \quad \forall t \le \tau \le t+h\ , \quad \forall \mathbf{z},\mathbf{w} \in K \subset D\ , \tag{2.2.29}$$

with compact set $K \subset D$, for which we assume (retrospectively) that $\mathbf{y}_h(\tau) \in K$ for all $t \leq \tau \leq t+h$. Then, for all $\mathbf{y}, \mathbf{z} \in (C^0([t, t+h]))^d$, $\mathbf{y}(\tau), \mathbf{z}(\tau) \in K \; \forall t \leq \tau \leq t+h$,

$$\|F(\mathbf{y}(\cdot)) - F(\mathbf{z}(\cdot))\|_\infty \leq h \, \|\mathsf{P}(\mathbf{f}(\cdot., \mathbf{y}(\cdot)) - \mathbf{f}(\cdot, \mathbf{z}(\cdot)))\|_\infty \overset{(2.2.29)}{\leq} h \, \|\mathsf{P}\| \, L \, \|\mathbf{y}(\cdot) - \mathbf{z}(\cdot)\| \; .$$

$$\boxed{|h| < \frac{1}{\|\mathsf{P}\| L}} \quad \Rightarrow \quad F \text{ is contraction.}$$

For sufficiently small $|h|$, $F(\mathbf{y}(\cdot))$ stays in the neighborhood of the constant function $\mathbf{y}_0$, if $\mathbf{y}(\cdot)$ is chosen from there.

Therefore, the assumptions of the fixed-point theorem 2.2.9 are satisfied. □

The projection property of P has not been used in the proof.

✎ Notation: $\tau \mapsto \mathbf{y}(\tau) \mathrel{\hat{=}}$ solution of the IVP $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$, $\mathbf{y}(t) = \mathbf{y}_0 \in D$,

$\tau \mapsto \mathbf{y}_h(\tau) \mathrel{\hat{=}}$ solution of (2.2.24) (for initial value $\mathbf{y}_0$)

Theorem 2.2.30 (One step error estimate for single step projection method)**.**
Let $\mathbf{f}$ *satisfy a local Lipschitz condition (→ Def. 1.3.2). Then, there exists* $h_0 > 0$*, such that*

$$\exists C > 0: \quad \|\mathbf{y} - \mathbf{y}_h\|_\infty \leq Ch\|(Id - \mathsf{P})(\mathbf{f}(\cdot, \mathbf{y}(\cdot)))\|_\infty \quad \forall |h| \leq h_0 \ .$$

projection error !

Thm. 2.2.30 $\Rightarrow$ consistency error estimate for single step projection method:

$$\|\boldsymbol{\tau}(t, \mathbf{y}, h)\| := \left\|\boldsymbol{\Phi}^{t,t+h}\mathbf{y} - \boldsymbol{\Psi}^{t,t+h}\mathbf{y}\right\| \leq Ch\,\|(Id - \mathsf{P})(\mathbf{f}(\cdot, \mathbf{y}(\cdot)))\|_\infty \ ,$$

with $C > 0$ independent of

- $\mathbf{y} \in K$, $K \hat{=}$ compact neighborhood of solution trajectory $t \mapsto \mathbf{y}(t)$, $t_0 \leq t \leq T$,
- sufficiently small time step $h > 0$.

Generalization of Lemma 2.2.13:

Lemma 2.2.31 (Lipschitz continuity of the increment function of single step projection methods)**.**

Let $\mathbf{f}$ *satisfy a local Lipschitz condition (→ Def. 1.3.2). Then, for every* $(t, \mathbf{y}) \in \Omega$*, there exists a* h_0 *such that, for* $|h| < h_0$*,*

$$\boldsymbol{\Psi}^{t,t+h}\mathbf{y} = \mathbf{y} + h\boldsymbol{\psi}(t, \mathbf{y}, h) \ ,$$

with an increment function $\boldsymbol{\psi}$ *(→ Lemma 2.1.9) that is locally Lipschitz continuous (→ Def. 1.3.2) w.r.t. the state variable* $\mathbf{y}$*.*

Proof. Use compactness arguments, cmp. proof of Thm. 2.1.19. WLOG assume global Lipschitz condition

$$\exists L > 0: \quad \|\mathbf{f}(\tau, \mathbf{z}) - \mathbf{f}(\tau, \mathbf{w})\| \leq L \|\mathbf{z} - \mathbf{w}\| \quad \forall \mathbf{z}, \mathbf{w} \in D, t \leq \tau \leq t + h \ .$$

Like in the proof of Lemma 2.2.27: if $\mathbf{y}_h, \mathbf{z}_h \in (C^0([t, t+h]))^d$ solutions of (2.2.26) for "initial values" $\mathbf{y}_0, \mathbf{z}_0 \in D$, there holds

$$|h| < \frac{1}{hL\,\|\mathsf{P}\|} \quad \Rightarrow \quad \|\mathbf{y}_h - \mathbf{z}_h\|_\infty \leq \frac{1}{1 - hL\,\|\mathsf{P}\|} \|\mathbf{y}_0 - \mathbf{z}_0\| \ . \tag{2.2.32}$$

guarantees existence of solutions of (2.2.26)

$$\text{(2.2.26)} \quad \Rightarrow \quad \boldsymbol{\Psi}^{t,t+h}\mathbf{y}_0 = \mathbf{y}_0 + \underbrace{\int_t^{t+h} \mathsf{P}(\mathbf{f}(\cdot,\mathbf{y}_h(\cdot)))(\xi)\,\mathrm{d}\xi}_{=h\boldsymbol{\psi}(t,\mathbf{y}_0,h)} =: \mathbf{y}_0 + h\boldsymbol{\psi}(t,\mathbf{y}_0,h)\ .$$

$$\begin{aligned} |\boldsymbol{\psi}(t,\mathbf{y}_0,h) - \boldsymbol{\psi}(t,\mathbf{z}_0,h)| &\leq \frac{t}{t+h}\int_0^h \mathsf{P}(\mathbf{f}(\cdot,\mathbf{y}_h(\cdot)) - \mathbf{f}(\cdot,\mathbf{z}_h(\cdot)))(\xi)\,\mathrm{d}\xi \\ &\leq \|\mathsf{P}\|\,L\,\|\mathbf{y}_h - \mathbf{z}_h\|_\infty \overset{\text{(2.2.32)}}{\leq} \frac{\|\mathsf{P}\|\,L}{1-hL\,\|\mathsf{P}\|}\,\|\mathbf{y}_0 - \mathbf{z}_0\|\ . \quad \square \end{aligned}$$

$(\mathbf{y}_k)_{k=0}^N \mathrel{\hat{=}}$ grid function, generated by single step projection method for $\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y})$ on time grid $\{t_0 < t_1 < \cdots < t_N = T\} \subset J(t_0,\mathbf{y}_0)$: $\mathbf{y}_{k+1} = \boldsymbol{\Psi}^{t_k,t_{k+1}}\mathbf{y}_k$

with Lemma 2.2.31 & proof of Thm. 2.1.19 (discrete Gronwall Lemma 2.1.20):

Global error estimate for single step projection method (on time grid)

$$\|\mathbf{y}_k - \mathbf{y}(t_k)\| \leq C \max_{j=1,\dots,N} \|(Id - \mathsf{P})(\mathbf{f}(\cdot,\mathbf{y}(\cdot)))\|_{\infty,[t_{j-1},t_j]} \frac{\exp(L(h_1+\cdots+h_k)) - 1}{L}\ , \tag{2.2.33}$$

for $k = 1,\dots,N$, h_j sufficiently small, $C > 0$ independent of h_j, k.

2.2.3 Convergence of collocation method

2.2.3.1 Order of consistency

Question: (Order of)consistency ($\rightarrow$ convergence, Sect. 2.1.3) of collocation methods **?**

Rem. 2.2.21 ➢ consequence of Lemma 2.2.31:

Lemma 2.2.34 (Consistency of collocation methods)**.**
Under the assumptions of Lemma 2.2.7, every collocation single step method is consistent ($\rightarrow$ Def. 2.1.8).

Reminder: Defining equations of a single step collocation method

$$\mathbf{y}_h(t_1) = \mathbf{y}_0 + h\sum_{i=1}^{s} b_i \mathbf{k}_i \;, \qquad \mathbf{k}_i = \mathbf{f}(t_0 + c_i h, \mathbf{y}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j) \;. \quad \text{with} \quad a_{ij} = \int_0^{c_i} L_j(\tau)\,\mathrm{d}\tau \;, \quad b_i = \int_0^1 L_i(\tau)\,\mathrm{d}\tau \;. \tag{2.2.3}$$

✎ Notation: $t \in [t_0, t_0 + h] \mapsto \mathbf{y}_h(t) \hat{=}$ approximate solution generated by collocation method, polynomial of degree s, compare (2.2.1)

➢ (one step) error function $\mathbf{e}(t) := \mathbf{y}(t) - \mathbf{y}_h(t), \quad t_0 \le t \le t_1$

Lemma 2.2.36 ((Suboptimal) order of consistency of collocation methods)**.**
The single step collocation method (2.2.3) *for* $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$ *is consistent of order* s *(→ Def. 2.1.13) for sufficiently smooth right hand side* $\mathbf{f}$*.*

Auxiliary tool for the proof: Estimate the remainder in the polynomial interpolation ($\rightarrow$ course "Numerical methods")

Lemma 2.2.37 (Estimation of the remainder in the polynomial interpolation). $\rightarrow$ *[13, Satz 7.16]*

Let $f \in C^{n+1}([x_0, x_n])$, $x_0 < x_1 < \ldots < x_n$, and let $p \in \mathcal{P}_n$ be the interpolation polynomial of f for points x_i (i.e., $p(x_i) = f(x_i)$). Then, there holds

$$|f^{(k)}(x) - p^{(k)}(x)| \leq \frac{|x_n - x_0|^{n+1-k}}{(n+1-k)!} \max_{x_0<\xi<x_n} |f^{(n+1)}(\xi)| \quad \forall x_0 \leq x \leq x_n,\ k = 0, \ldots, n+1 .$$

P. Grohs

rev 63606, February 17, 2014

Proof of Lemma 2.2.35. Due to Lemma 2.2.36, there holds the following estimate for the interpolation error for P_{s-1} on $[t, t+h]$:

$$\|(Id - \mathsf{P}_{s-1})\mathbf{f}(\cdot, \mathbf{y}(\cdot))\|_\infty \leq h^s \left\| \frac{d^s}{dt^s}\mathbf{f}(\cdot, \mathbf{y}(\cdot)) \right\|_\infty . \tag{2.2.38}$$

Based on the assumption of sufficient smoothness of $\mathbf{f}$, which is transferred to the exact solution $\mathbf{y}$ of the IVP, the right hand side in (2.2.37) is asymptotically $O(h^s)$ for $h \to 0$.

Together with Thm. 2.2.30, this implies the bound $O(h^{s+1})$ for the one step error (= consistency error). □

[Beitrag von J. Mitrovic, FS 2011]

In the above proof, we actually deduced a bound for the consistency error:
$$||\tau(t_0, y_0, h)|| \leq ||\mathbf{y}_h - \mathbf{y}||_\infty \leq Ch \underbrace{||(\mathbf{Id} - P)\mathbf{f}(\cdot, \mathbf{y}(\cdot))||_\infty}_{\text{projection error}} \quad \forall t_0 \leq t \leq t_0 + h..$$

"Direct" proof of Lemma 2.2.35. (for autonomous differential equations $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $t_0 = 0$)

- $t \mapsto \mathbf{y}(t) \hat{=}$ exact solution of IVP $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\mathbf{y}(0) = \mathbf{y}_0$
- $t \mapsto \mathbf{y}_h(t)$, $0 \leq t \leq h$, polynomial, approximate solution from one step of the collocation method, $\mathbf{y}_h(0) = \mathbf{y}_0$.

(Assumption: h sufficiently small, compare Lemma 2.2.7, $h \in J(\mathbf{y}_0)$)

Again, a simplifying assumption: $\mathbf{f}$ globally Lipschitz continuous

$$\exists L > 0: \quad \|\mathbf{f}(\mathbf{z}) - \mathbf{f}(\mathbf{w})\| \leq L \|\mathbf{z} - \mathbf{w}\| \quad \forall \mathbf{z}, \mathbf{w} \in D \,. \tag{2.2.39}$$

To examine consistency, consider the one step error function

$$\mathbf{e}(t) := \mathbf{y}(t) - \mathbf{y}_h(t) \quad \succ \quad \boldsymbol{\tau}(\mathbf{y}_0, h) = \mathbf{e}(h) \; . \tag{2.2.40}$$

From the collocation conditions (2.2.1):

$$\dot{\mathbf{y}}_h(t) = \sum_{i=1}^{s} \mathbf{f}(\mathbf{y}_h(c_i h)) \cdot L_i(\tau) \; , \quad \tau := \frac{t}{h} \; . \tag{2.2.41}$$

Based on the solution property of $t \mapsto \mathbf{y}(t)$

$$\dot{\mathbf{y}}(t) = \mathbf{f}(\mathbf{y}(t)) = \sum_{i=1}^{s} \mathbf{f}(\mathbf{y}(c_i h)) \cdot L_i(\tau) + \mathbf{r}(t) \; , \quad \tau := \frac{t}{h} \; . \tag{2.2.42}$$

Interpolation polynomial $\in \mathcal{P}_{s-1}$ for function $t \mapsto \mathbf{f}(\mathbf{y}(t))$ and nodes $c_i h$ on $[0, h]$

$\mathbf{r} \mathrel{\hat{=}}$ Remainder for polynomial interpolation, compare Lemma 2.2.36, satisfies

$$\left\| \mathbf{r}^{(k)}(t) \right\| \leq \underbrace{\frac{1}{(s-k)!} \max_{0<t<T} \left\| \mathbf{y}^{(s+1)}(t) \right\|}_{\text{independent of } h} h^{s-k} \leq C h^{s-k} \; , \quad k = 0, \dots, s \; . \tag{2.2.43}$$

Convention: All constants C are independent of (sufficiently small) h, may depend on $\mathbf{y}(t)$, $\mathbf{f}$, parameters of the collocation method, etc.

From (2.2.40) & (2.2.41) & integration ➢ expression of one step error function

$$\boxed{\mathbf{e}(t) = h\sum_{i=1}^{s}\Delta\mathbf{f}(c_i h)\cdot\int_0^{\tau} L_i(\sigma)\,\mathrm{d}\sigma + \int_0^t \mathbf{r}(\sigma)\,\mathrm{d}\sigma\ ,\quad 0\le\tau\le 1}\ , \tag{2.2.44}$$

with $\Delta\mathbf{f}(t) = \mathbf{f}(\mathbf{y}(t)) - \mathbf{f}(\mathbf{y}_h(t))$.

$$\text{(2.2.38)} \quad\Rightarrow\quad \|\Delta\mathbf{f}(t)\| \le L\,\|\mathbf{e}(t)\|\ . \tag{2.2.45}$$

$$\overset{\text{(2.2.43) \& (2.2.42)}}{\Longrightarrow}\quad \|\mathbf{e}\|_\infty := \max_{0\le t\le h}\|\mathbf{e}(t)\| \le C_1 L h\,\|\mathbf{e}\|_\infty + C_2 h^{s+1}\ , \tag{2.2.46}$$

with constants $C_1, C_2 > 0$ that are *independent of* h (, but depend on the collocation points c_i and $\mathbf{y}$).

$$C_1 L h_0 < 1 \quad\Rightarrow\quad \|\boldsymbol{\tau}(\mathbf{y},h)\| \le \|\mathbf{e}\|_\infty \le \frac{C_2}{1-C_1 h_0 L}h^{s+1} \quad \forall |h|\le h_0\ . \tag{2.2.47}$$

☞ We conclude: The collocation method has at least the order of consistency s. □

From (2.2.43) and (2.2.42), we can even deduce that, for $k = 0,\dots,s$,

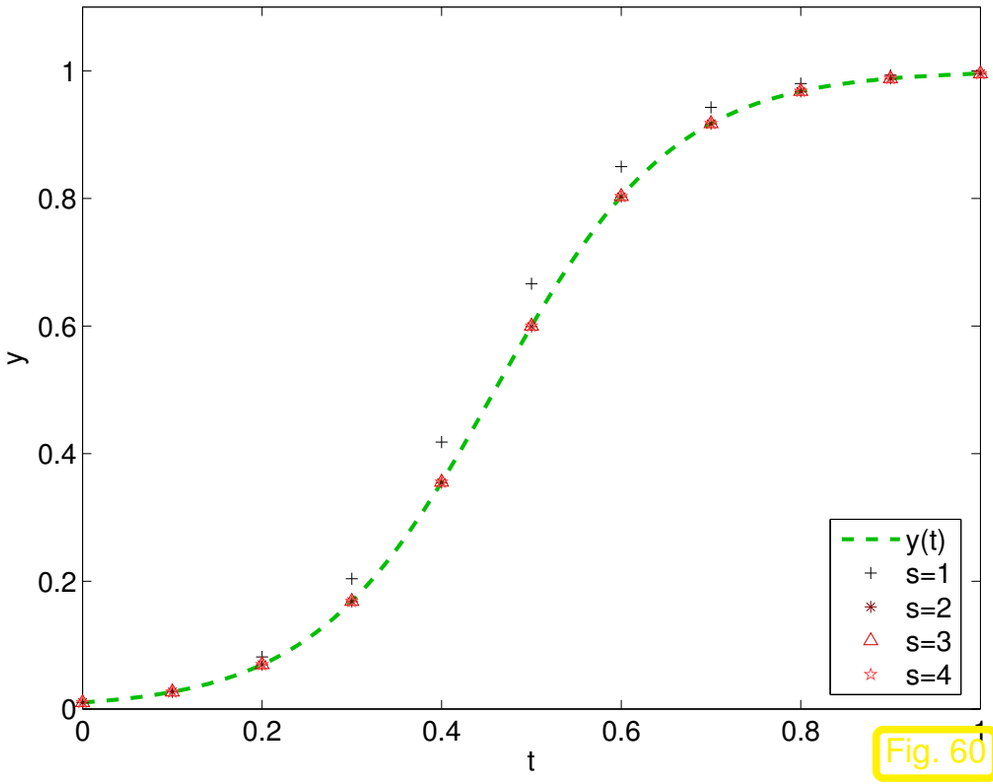
$$\max_{0\le t\le h}\left\|\mathbf{e}^{(k)}(t)\right\| \le C_1(k) L h\,\|\mathbf{e}\|_\infty + C_2(k) h^{s+1-k}$$

$$\Rightarrow\quad \max_{0\le t\le h}\left\|\mathbf{e}^{(k)}(t)\right\| \le C(k) h^{s+1-k}\ , \tag{2.2.48}$$

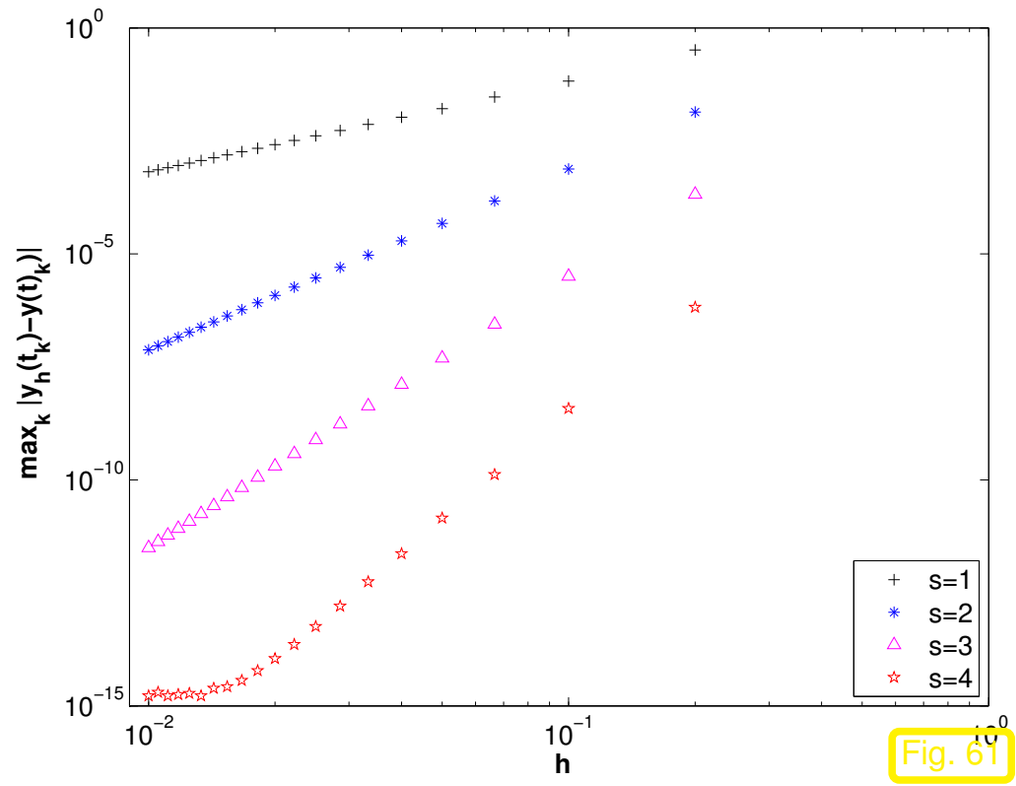
with constants $C_1(k), C_2(k), C(k) > 0$ *independent of* h.

Example 2.2.49 (Convergence of single step Gauss collocation method). → Ex. 2.2.17

- Scalar logistic differential equation (1.2.2), $\lambda = 10$, $y(0) = 0.01$, $T = 1$
- Single step Gauss collocation method (2.2.3) for $s = 1, \dots, 4$, uniform time step h



$y_h(j/10)$: Gauss collocation method



Convergence of error $\max_k |y_k - y(t_k)|$

Numerical rates of convergence :
(computed by linear regression)

$$s = 1 \;:\; p = 1.96$$
$$s = 2 \;:\; p = 4.01$$
$$s = 3 \;:\; p = 6.00$$
$$s = 4 \;:\; p = 8.02$$

Observation: rates of convergence are twice as big as the lower bound from Lemma 2.2.35!

Comparison of (empirical) rates of convergence ▷

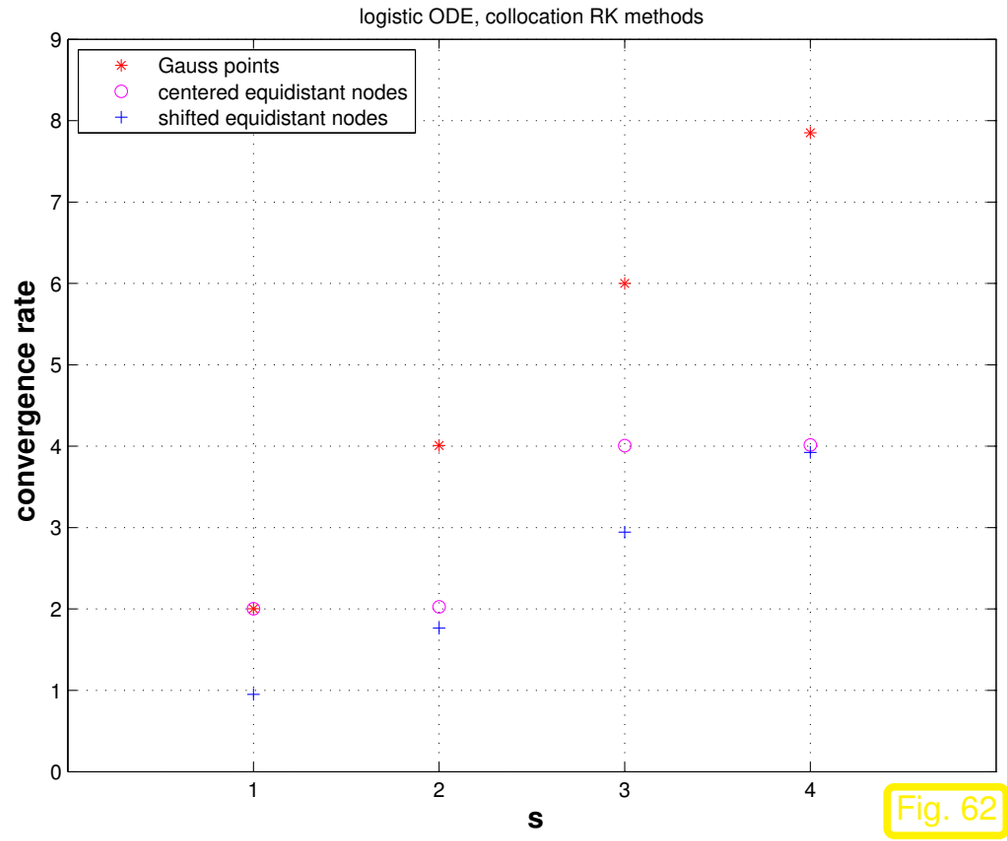


Fig. 62

◇

Note: Collocation method for $\dot{\mathbf{y}} = f(t, \mathbf{y})$ with (relative) collocation points $c_i \in [0,1]$, $i = 1, \dots, s$, $s \in \mathbb{N}$ ➢ coefficients b_i, a_{ij} in (2.2.3).

▶ Assigned quadrature formula, Rem. 2.2.18: $$Q(f) = h \sum_{i=1}^{s} b_j f(t_0 + c_j h) \ . \quad \text{(2.2.50)}$$

Ex. 2.2.48 implies the conjecture that the order of consistency of a collocation method coincides with the order of the quadrature formula that has been assigned according to (2.2.49) .

Theorem 2.2.51 (Order of consistency of collocation methods)**.**
The order of consistency (→ Def. 2.1.13) of a single step collocation method coincides with the order of the assigned quadrature formula.

Auxiliary tool for the proof: error estimate for numerical quadrature (→ course "Numerical methods")

s point quadrature formula on $[a,b]$:

$$Q(f) := (b-a)\sum_{i=1}^{s} b_i f(a + c_i(b-a)) \approx \int_a^b f(x)\,\mathrm{d}x \,. \tag{2.2.52}$$

Assumption: inner nodes $0 \leq c_i \leq 1$, $i = 1, \dots, s$

$\rightarrow$ Quadrature formula of order $n+1$

Lemma 2.2.53 (Estimating the quadrature error). *If a quadrature formula* (2.2.51) *is exact for polynomials of degree* $\leq n$*, then there holds*

$$f \in C^{n+1}([a,b]) \quad \Rightarrow \quad \left| Q(f) - \int_a^b f(x)\,\mathrm{d}x \right| \leq C \frac{(b-a)^{n+2}}{(n+1)!} \max_{a<x<b} |f^{(n+1)}(x)| \,,$$

with $C = 1 + \sum_{i=1}^{s} |b_i|$.

P. Grohs

rev 63606, February 17, 2014

Assumption: $\mathbf{f}$ "'sufficiently" smooth, locally Lipschitz continuous ($\rightarrow$ Def. 1.3.2)

Proof of Thm. 2.2.50 (for autonomous differential equation $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $p \mathrel{\hat{=}}$ order of the underlying quadrature formula)

Idea: Interpret $t \mapsto \mathbf{y}_h(t)$ as solution of a perturbed initial value problem !

$$\dot{\mathbf{y}}_h(t) = \mathbf{f}(\mathbf{y}_h(t)) + \underbrace{\dot{\mathbf{y}}_h(t) - \mathbf{f}(\mathbf{y}_h(t))}_{\delta(t)} \,, \quad 0 \leq t \leq h \,. \tag{2.2.54}$$

Due to $\dot{\mathbf{y}}(t) = \mathbf{f}(\mathbf{y}(t))$, there holds for the one step error function $\mathbf{e}(t) = \mathbf{y}(t) - \mathbf{y}_h(t)$

$$\dot{\mathbf{e}}(t) = \mathbf{f}(\mathbf{y}(t)) - \mathbf{f}(\mathbf{y}_h(t)) - \delta(t) \ , \quad 0 \le t \le h \ .$$

Auxiliary tool: Taylor expansion

$$\varphi(1) - \varphi(0) = \varphi'(0) + \int_0^1 (1-\tau)\varphi''(\tau)\,\mathrm{d}\tau \ ,$$

for $\varphi(\xi) := \mathbf{f}(\mathbf{y}(t) + \xi(\mathbf{y}_h(t) - \mathbf{y}(t)))$ with chain rule: ($' \hat{=}$ derivative w.r.t. ξ)

$$\begin{aligned} \varphi'(\xi) &= D\mathbf{f}(\mathbf{y}(t) + \xi(\mathbf{y}_h(t) - \mathbf{y}(t))) \cdot (\mathbf{y}_h(t) - \mathbf{y}(t)) \ , \\ \varphi''(\xi) &= D^2\mathbf{f}(\mathbf{y}(t) + \xi(\mathbf{y}_h(t) - \mathbf{y}(t)))\,(\mathbf{y}_h(t) - \mathbf{y}(t), \mathbf{y}_h(t) - \mathbf{y}(t)) \ . \end{aligned}$$

Substitution in the formula for the one step error function:

$$\dot{\mathbf{e}}(t) = \varphi(0) - \varphi(1) - \delta(t) = D\mathbf{f}(\mathbf{y}(t))\mathbf{e}(t) - \underbrace{\int_0^1 (1-\tau) D^2\mathbf{f}(\mathbf{y}(t) + \tau\mathbf{e}(t))(\mathbf{e}(t), \mathbf{e}(t))\mathrm{d}\tau}_{=:\rho(t)} - \delta(t) \ .$$

Hereby, we have the obvious estimate:

$$\|\rho(t)\| \le \max_{\mathbf{y}\in K} \left\| D^2\mathbf{f}(\mathbf{y}) \right\| \cdot \|\mathbf{e}(t)\|^2 \overset{\text{Lemma 2.2.35}}{\le} Ch^{2s+2} \ , \tag{2.2.55}$$

where $K = \{\mathbf{z} \in D\colon \|\mathbf{z} - \mathbf{y}(t)\| \le R\}$ with $R > 0$ and $C > 0$ independent of t *independent of* h.

➢ the one step error function solves the initial value problem

$$\boxed{\dot{\mathbf{e}} = D\mathbf{f}(\mathbf{y}(t))\mathbf{e} - \rho(t) - \delta(t) \quad , \quad \mathbf{e}(0) = 0} \; . \tag{2.2.56}$$

If we consider $\rho(t), \delta(t)$ as functions of t, then (2.2.55) is a *non-autonomous linear* differential equation.

➢ solution by general variation-of-constant formula (1.3.18):

$$\mathbf{e}(t) = -\int_0^t \mathbf{W}(t;\mathbf{y}_0)\mathbf{W}(\tau;\mathbf{y}_0)^{-1}(\rho(\tau) + \delta(\tau))\,\mathrm{d}\tau \; , \quad 0 \le t \le h \; .$$

with propagation matrix $t \mapsto \mathbf{W}(t;\mathbf{y}_0)$, cmp. (1.3.33), Sect. 1.3.3.4. It solves the initial value problem for the variational equation (1.3.34)

$$\dot{\mathbf{W}}(t;\mathbf{y}_0) = D\mathbf{f}(\mathbf{y}(t))\mathbf{W}(t;\mathbf{y}_0) \quad , \quad \mathbf{W}(0;\mathbf{y}_0) = \mathbf{I} \; .$$

The propagation matrix is of course independent of h; therefore

$$\exists C > 0 \quad \text{independent of } h\text{:} \quad \left\| \mathbf{W}(t;\mathbf{y}_0)\mathbf{W}(\tau;\mathbf{y}_0)^{-1} \right\| \le C \quad \forall 0 \le t, \tau \le h \; .$$

$$\overset{(2.2.54)}{\Rightarrow} \quad \left\| \int_0^t \mathbf{W}(t;\mathbf{y}_0)\mathbf{W}(\tau;\mathbf{y}_0)^{-1}\rho(\tau)\,\mathrm{d}\tau \right\| \le Ch^{2s+3} \; ,$$

with $C > 0$ independent of h.

Note: (2.2.39) ➢ consistency error determined by $\mathbf{e}(h)$!

Ingenious idea: estimation of $\int_0^h \mathbf{W}(t;\mathbf{y}_0)\mathbf{W}(\tau;\mathbf{y}_0)^{-1}\delta(\tau)\,\mathrm{d}\tau$ as quadrature error **!**

Due to $\delta(c_i h) = 0$ (collocation condition (2.2.1) !):

$$\underbrace{\sum_{i=1}^{s} b_i \mathbf{W}(t;\mathbf{y}_0)\mathbf{W}(c_i h;\mathbf{y}_0)^{-1}\underbrace{\delta(c_i\tau)}_{=0}}_{\text{Quadrature formula on } [0,h] \text{ for } \tau\mapsto\mathbf{W}(t;\mathbf{y}_0)\mathbf{W}(\tau;\mathbf{y}_0)^{-1}\delta(\tau)} = 0 \quad \forall 0 \le t \le h\,.$$

➢ With the estimate for the quadrature error from Lemma 2.2.52

$$\left\| \int_0^h \mathbf{W}(t;\mathbf{y}_0)\mathbf{W}(\tau;\mathbf{y}_0)^{-1}\delta(\tau)\,\mathrm{d}\tau - \sum_{i=1}^{s} b_i \mathbf{W}(t;\mathbf{y}_0)\mathbf{W}(c_i h;\mathbf{y}_0)^{-1}\delta(c_i\tau) \right\| \le C_3 h^{p+1}\,,$$

with

$$C_3 := \frac{1}{p!}\max_{0\le\tau\le h}\left\| \frac{d^p}{d\tau^p}\left\{\tau \mapsto \mathbf{W}(h;\mathbf{y}_0)\mathbf{W}(\tau;\mathbf{y}_0)^{-1}\delta(\tau)\right\}\right\|\,.$$

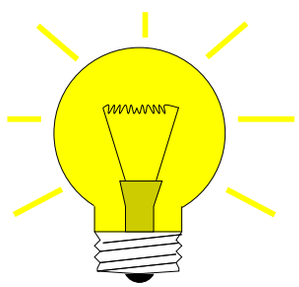
In contrast to $\mathbf{W}(t;\mathbf{y})$, $\delta(t) := \dot{\mathbf{y}}_h(t) - \mathbf{f}(\mathbf{y}_h(t))$ depends on h. However, thanks to the bounds from (2.2.47), all derivatives of $\mathbf{y}_h$ are uniformly bounded in h **!**. Therefore, we can also bound C_3 independently of h.

$$\blacktriangleright \quad \left\| \int_0^h \mathbf{W}(h;\mathbf{y}_0)\mathbf{W}(\tau;\mathbf{y}_0)^{-1}\delta(\tau)\,\mathrm{d}\tau \right\| \le C h^{p+1}\,.$$

The claim follows from the estimate for the ρ term, cmp. Def. 2.1.13 □

► s-step implicit single step Gauss collocation method are of the order $2s$

2.2.3.2 Spectral convergence



Reminder: Polynomial interpolation: "degree $\to \infty$ $\Rightarrow$ error $\to 0$"
(for suitable nodes, e.g., Tschebyscheff nodes [13, Sect. 7.1.4])

► For single step Gauss collocation method:

convergence for $s \to \infty$ **?**

Example 2.2.57 (Convergence of <u>global</u> Gauss collocation method).

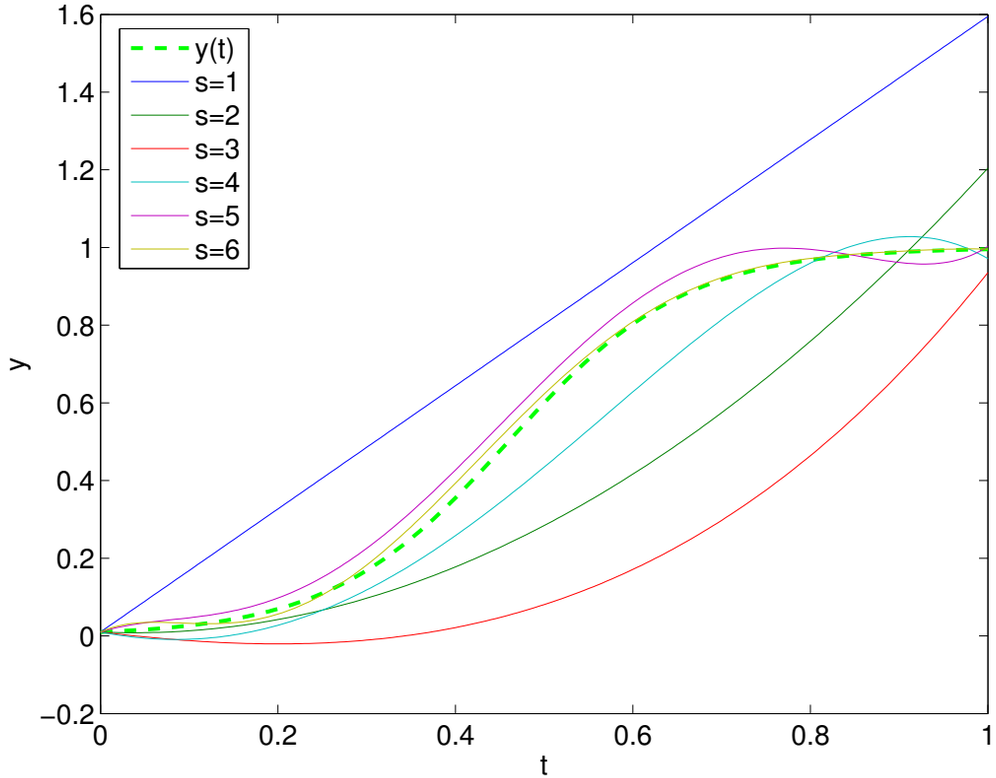
This example studies the *one step error* of collocation methods in dependence of the number of collocation points $\leftrightarrow$ polynomial degree s. So far we only considered the strategy to refine the time grid in order to achieve a more accurate solution.

Logistic differential equation (→ Ex. 1.2.1)

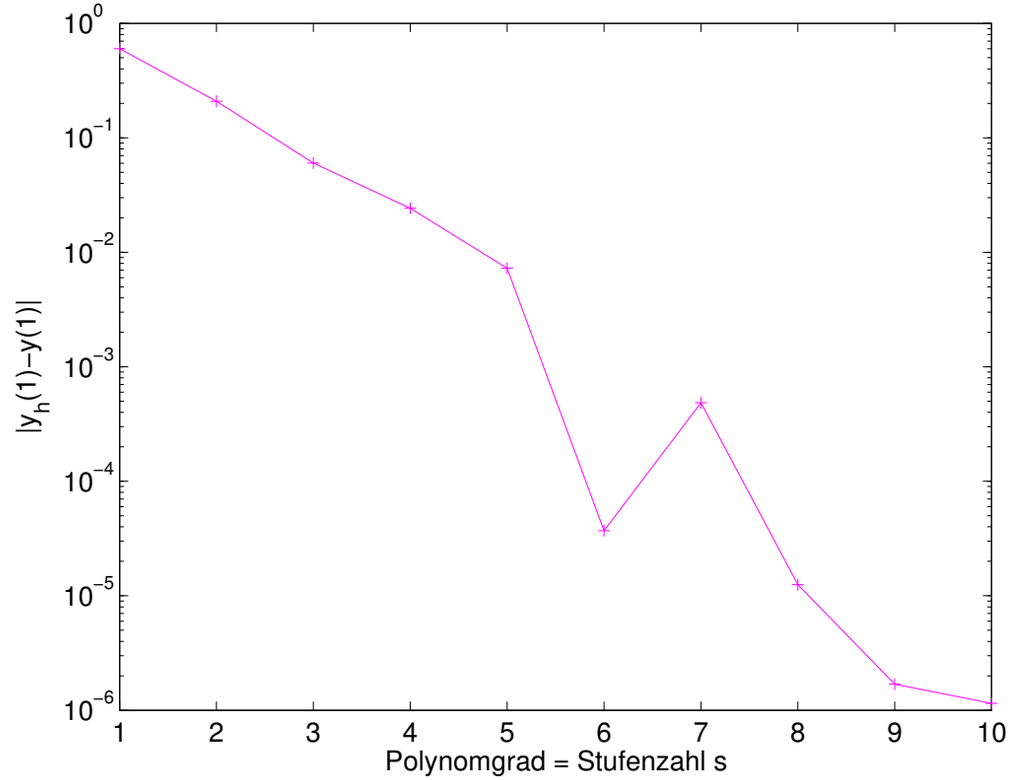
$$\dot{y} = \lambda y(1-y)\ , \quad y_0 \in]0,1[\quad \Rightarrow \quad y(t) = \frac{1}{1+(y_0^{-1}-1)e^{-\lambda t}}\ , \quad t \in \mathbb{R}\ . \tag{2.2.58}$$

Numerical experiments with Gauss collocation method on $[0,1]$, $y_0 = 0.01$, $\lambda = 10$:
(Solution of equations for increments $\mathbf{k}_i$: MATLAB `fsolve`, tolerance 10^{-9})

Here: collocation method as global integration method



Approximate solutions $y_h(t)$



▷ Exponential convergence in s (Why ?)

◇

P. Grohs

rev 63606, February 17, 2014

Related: convergence analysis based on Thm. 2.2.30

☞ Required: Estimate of spectral interpolation error for polynomial interpolation in Gauss nodes, compare map **??** (spectral: error estimates dependent on polynomial degree, a new aspect compared to course “Numerical methods”).

Estimating the error for polynomial interpolation of analytic functions

Example 2.2.59 (Interpolation error for polynomial interpolation at Gauss nodes).

Interpolant: solution of the logistic differential equation (1.2.2) on $[-1,1]$, cmp. Ex. 1.2.1:

$$y(t) = \frac{1}{1+\exp(-\frac{1}{2}\lambda t)} \; .$$

Error:

$$\mathrm{err} := \max_{-1\leq t\leq 1} |y(t) - p_n(t)| \; ,$$

$p_n(t) \mathrel{\hat{=}}$ interpolation polynomial on $y(t)$, degree $n-1$, for n Gauss nodes.

Approximate evaluation of err by scanning on very fine grid

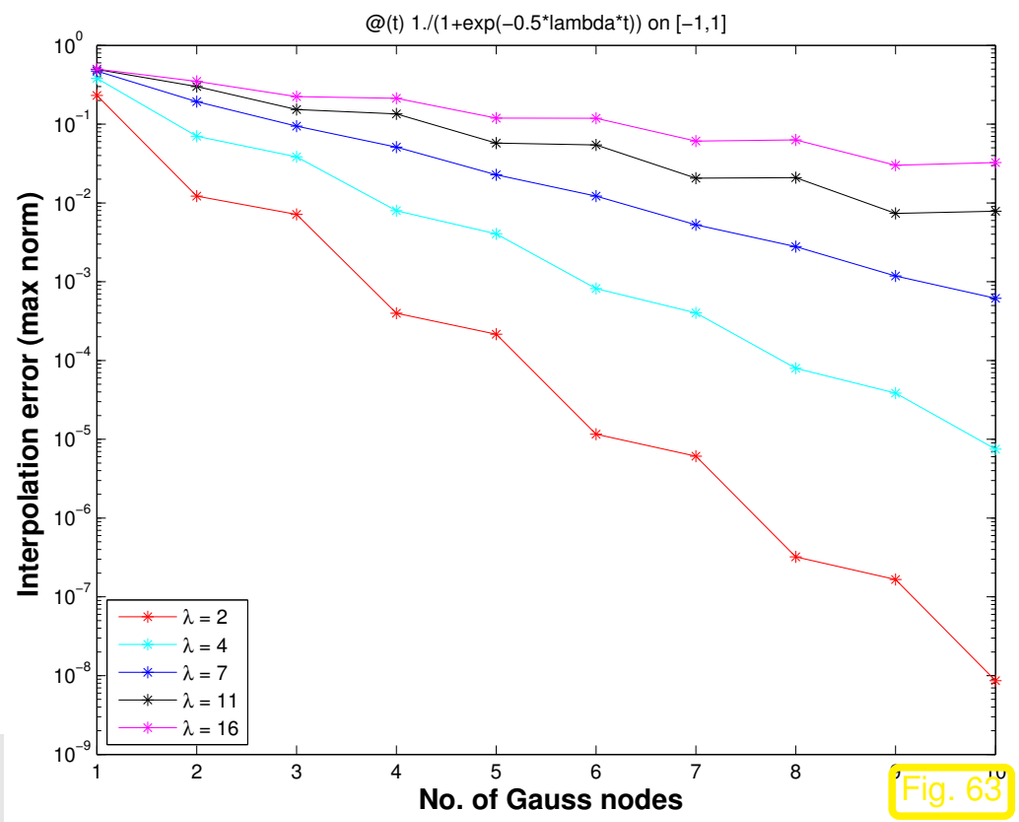


Fig. 63

Listing 2.1: Computation of error of polynomial interpolation in maximum norm

```
function errinf = polyintperr(fun,nodes,span)
% Error of polynomial interpolation in maximum norm
% fun :  handle to function to be interpolated
% nodes :  interpolation nodes
% span :  evaluation interval (default [-1,1])

if (nargin < 3), span = [-1,1];  end
n = length(nodes);

fval = zeros(1,n);
for j=1:n, fval(j) = fun(nodes(j)); end

p = polyfit(nodes,fval,n-1); % built-in polynomial interpolation

% Compute maximum norm by sampling on fine mesh
N    = 1000*n;
t    = span(1) + (0:N)*(span(2)-span(1))/N;
pval = polyval(p,t);
fval = zeros(1,N+1);
for j=1:N+1, fval(j) = fun(t(j)); end
errinf = max(abs(pval-fval));
```

Listing 2.2: Computation of interpolation error at Gauss nodes in maximum norm

```
function errinf = gaussintperr(fun,n)
% Error of polynomial interpolation in Gauss points in maximum norm
% fun :  handle to function to be interpolated
% n :  number of interpolation nodes

path(path,'../SupportScripts');
[nodes,weights] = GaussQuad(n);
errinf = polyintperr(fun,nodes');
```

Listing 2.3: Generation of plots for Ex. 2.2.58

```
function plotgaussintperr
% Error of Gaussian interpolation for solution of logistic differentialequation

rec = []; % Array for recording errors
k = 1;
for lambda=[2,4,7,11,16]
  sol = @(t) 1./(1+exp(-0.5*lambda*t));
  errs = [];
  for n=1:10, errs = [errs,gaussintperr(sol,n)]; end
  rec = [rec;errs];
  leg{k} = sprintf('\\lambda = %d',lambda);
  k = k+1;
end
```

```
figure('name','Gaussian interpolation error');
semilogy(1:10,rec(1,:),'r*-',...
      1:10,rec(2,:),'c*-',...
      1:10,rec(3,:),'b*-',...
      1:10,rec(4,:),'k*-',...
      1:10,rec(5,:),'m*-');
xlabel('{\bf No. of Gauss nodes}','fontsize',14);
ylabel('{\bf Interpolation error (max norm)}','fontsize',14);
title('@(t) 1./(1+exp(-0.5*lambda*t)) on [-1,1]');
legend(leg,'location','southwest');

print -depsc2 'gaussintperr.eps';
```

Observation: exponential convergence ($\rightarrow$ Def. 1.4.5) of the interpolation error, faster for smaller λ.

$\diamond$

Analysis of the interpolation error with auxiliary tools from *complex variables*:

The following should be known from complex variables: concept of a

- *holomorphic* function,
- Cauchy's integral theorem , and
- Laurent expansion.

Reminder, compare, e.g., [36, Ch. 13]:

Theorem 2.2.60 (Residue theorem)**.**
Let $D \subset \mathbb{C}$ be an open set, let $\Gamma \subset D$ be a simple closed integration path, and let $\Pi \subset D$ be a finite set.
For every in $D \setminus \Pi$ holomorphic *(analytic) function $f : D \setminus \Pi \mapsto \mathbb{C}$, there holds*

$$\frac{1}{2\pi i}\int_\gamma f(z)\,\mathrm{d}z = \sum_{p\in\Pi} \operatorname{res}_p f \ ,$$

where $\operatorname{res}_p f$ is the residue of f at point p.

Π often corresponds to the set of poles (points, for which $|f|$ has the "value" ∞) of f.

Definition 2.2.61 (Residue of a complex valued function)**.**
Let f be holomorphic in the punctured neighborhood of $p \in \mathbb{C}$. Then the residue $\operatorname{res}_p f$ of f at the point p is the coefficient a_{-1} of the Laurent expansion of f in p.

Sketch of proof. (of Thm. 2.2.59)

If f has in the punctured neighborhood of $p \in \mathbb{C}$ the convergent Laurent expansion

$$f(z) = \sum_{k=-\infty}^{\infty} a_k (z-p)^k,$$

then there holds for a (sufficiently small) circle γ around p

$$\frac{1}{2\pi i} \int_\gamma f(z)\,\mathrm{d}z = a_{-1}\ .$$

Then decompose $\int_\Gamma$ as indicated in the sketch and use Cauchy's integral theorem.

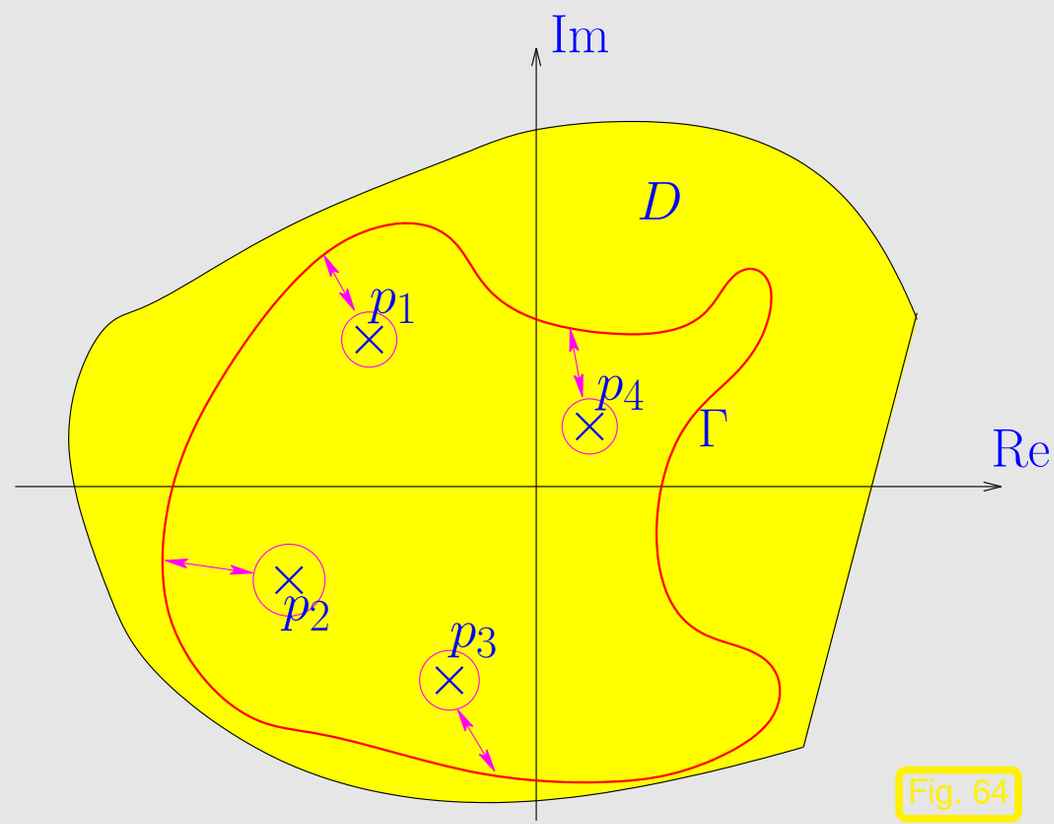


Fig. 64

P. Grohs

rev 63606, February 17, 2014

Lemma 2.2.62 (Residue formula for simple pole)**.**
If f is holomorphic in the punctured neighborhood of $p \in \mathbb{C}$ and $(z-p)f(z)$ is holomorphic in p, then there holds

$$\operatorname{res}_p f = \lim_{z \to p} (z-p) f(z) \ . \tag{2.2.63}$$

This directly implies:

Lemma 2.2.64 (Residue formula for quotients)**.**
If g, h are holomorphic in the neighborhood of $p \in \mathbb{C}$ and $h(p) = 0$, $h'(p) \neq 0$, then there holds

$$\operatorname{res}_p \frac{g}{h} = \frac{g(p)}{h'(p)} \ .$$

Consider: polynomial interpolation of $f \in C^0([a,b])$ in nodes $\tau_1 < \tau_2 < \ldots < \tau_s$, $s \in \mathbb{N}$

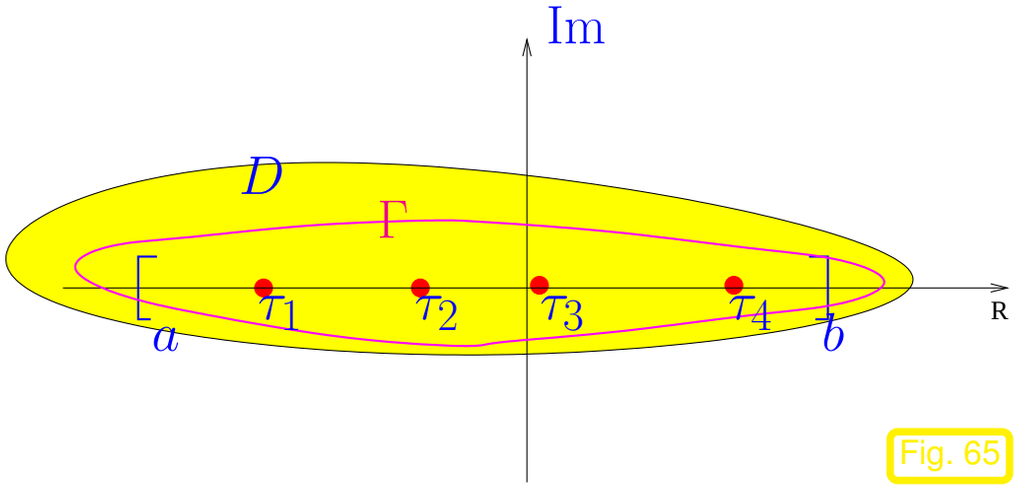


Fig. 65 ◁ Domain of analyticity (yellow)

Assumption 2.2.65 (Analyticity of the interpolant)**.**
f analytically continuable to the complex neighborhood $D \subset \mathbb{C}$ of $[a,b]$.

We consider the following function with the set of poles $\Pi = \{t, \tau_1, \dots, \tau_s\}$

$$g_t(z) := \frac{f(z)}{(z-t)P(z)}\,, \quad t \in [a,b] \setminus \{\tau_1, \dots, \tau_s\}\,, \quad P(z) := \alpha(z-\tau_1)\cdot\dots\cdot(z-\tau_s),\ \alpha \in \mathbb{R}\,.$$

➢ g_t is holomorphic on $D \setminus \{t, \tau_1, \dots, \tau_s\}$ (t is parameter!).)

▶ Application of residue theorem 2.2.59 on g_t with simple closed integration path $\Gamma \subset D$, which encloses $[a,b]$, compare the magenta curve in figure 65:

$$\frac{1}{2\pi i}\int_\Gamma g_t(z)\,\mathrm{d}z = \operatorname{res}_t g_t + \sum_{j=1}^{s} \operatorname{res}_{\tau_j} g_t \overset{\text{Lemma 2.2.63}}{=} \frac{f(t)}{P(t)} + \sum_{j=1}^{s} \frac{f(\tau_j)}{(\tau_j - t)P'(\tau_j)}$$

Possible as P has only *simple* roots !

P. Grohs

rev 63606, February 17, 2014

$$\boxed{f(t) = \underbrace{-\sum_{j=1}^{s} f(\tau_j) \underbrace{\frac{P(t)}{(\tau_j - t)P'(\tau_j)}}_{-\text{Lagrange polynomial !}}}_{\text{interpolation polynomial !}} + \underbrace{\frac{P(t)}{2\pi i}\int_\Gamma g_t(z)\,\mathrm{d}z}_{\text{interpolation error !}}} \; . \tag{2.2.66}$$

➢ Now to estimate: right hand side of

$$\boxed{|f(t) - \text{interpolation polynomial}(t)| \leq \left|\frac{P(t)}{2\pi i}\int_\Gamma \frac{f(z)}{(z-t)P(z)}\,\mathrm{d}z\right| \;, \quad a \leq t \leq b \; .} \tag{2.2.67}$$

P. Grohs

rev 63606, February 17, 2014

Blocks of the estimate:

- Upper bound for $|P(t)|$, $a \leq t \leq b$
- Lower bound for $|P(z)|$, $z \in \Gamma$ for a cleverly chosen integration path $\Gamma \subset D$

The bound in (2.2.66) obviously does not depend on α.

Estimates for Legendre polynomials

Reminder: for $\{\tau_j\}_{j=1}^{s} \hat{=}$ Gauss nodes in $[-1,1]$ ➤ $P(t) \hat{=} s^{\text{th}}$ Legendre polynomial (degree s)

✎ Notation: $P_n \hat{=}$ Legendre polynomial of degree $n \in \mathbb{N}_0$

Recursion formula: $$(n+1)P_{n+1}(t) - (2n+1)tP_n(t) + nP_{n-1}(t) = 0 \;, \tag{2.2.68}$$

Rodrigues formula: $$P_n(t) = \frac{1}{2^n n!}\frac{d^n}{dt^n}(t^2-1)^n \;. \tag{2.2.69}$$

(Start of recursion with $P_0 \equiv 1$, $P_1(t) = t$)

Legendre polynomials on $[-1,1]$ ▷

$$P_0(x) = 1 \;,$$
$$P_1(x) = x \;,$$
$$P_2(x) = 1/2(3x^2-1) \;,$$
$$P_3(x) = 1/2(5x^3-3x) \;,$$
$$P_4(x) = 1/8(35x^4-30x^2+3) \;,$$
$$P_5(x) = 1/8(63x^5-70x^3+15x) \;,$$
$$P_6(x) = 1/(16)(231x^6-315x^4+105x^2-5) \;.$$

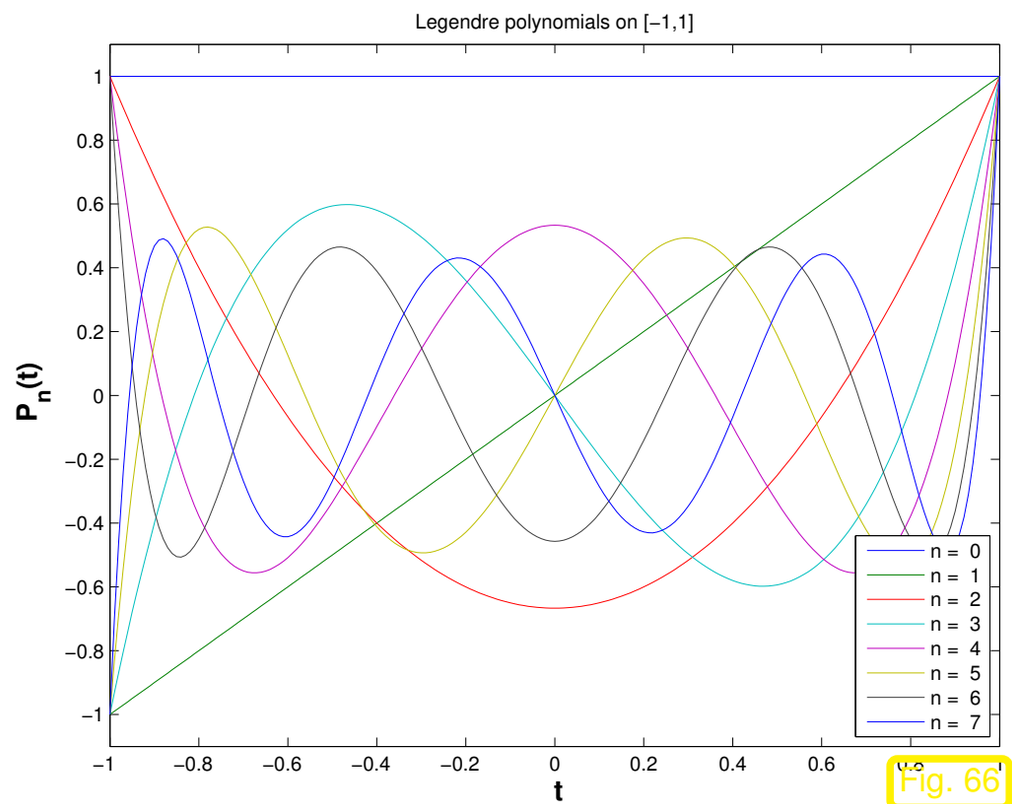


Fig. 66

P. Grohs

rev 63606, February 17, 2014

Conjecture: $|P_n(t)| \leq 1$ for all $-1 \leq t \leq 1$

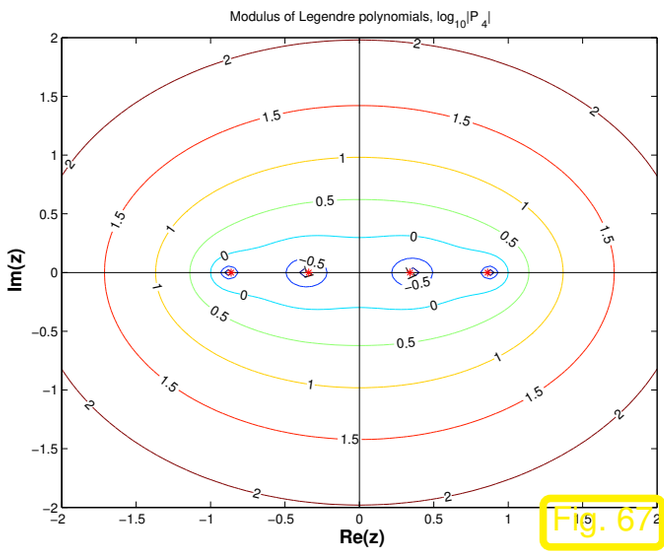


Fig. 67

level lines of $|P_4(z)|$

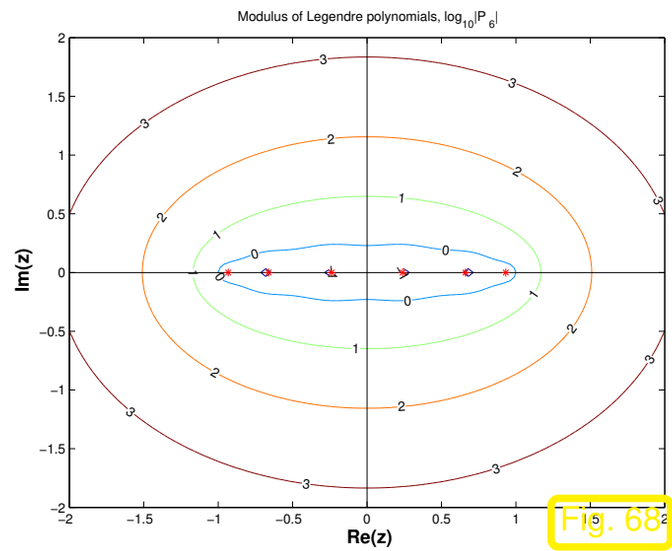


Fig. 68

level lines of $|P_6(z)|$

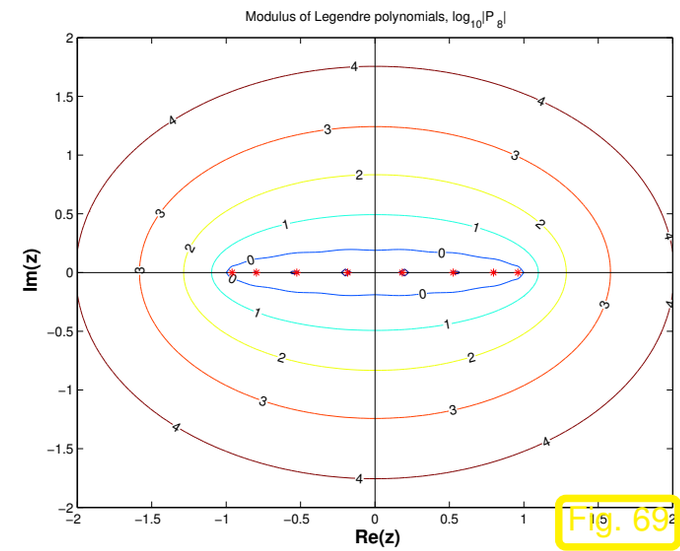


Fig. 69

level lines of $|P_8(z)|$

Observation/conjecture:

- Level lines of $|P_n(z)|$ are approximately *ellipses* with foci -1, 1.
- Exponential growth of $|P_n(z)|$ to extending ellipses with foci $\{-1, 1\}$.

Auxiliary tool for deducing upper and lower bounds for Legendre polynomials:

Generating function of the Legendre polynomial

$$\text{formal series:} \quad F_w(z) = \sum_{n=0}^{\infty} P_n(w) z^n \,, \quad z, w \in \mathbb{C} \,. \tag{2.2.70}$$

By differentiating term by term:

$$\begin{aligned} \frac{dF_w}{dz}(z) &= \sum_{n=0}^{\infty} (n+1) P_{n+1}(w) z^n \,, \\ \frac{d}{dz}(zF_w(z)) &= \sum_{n=0}^{\infty} (n+1) P_n(w) z^n \,, \\ z\frac{d}{dz}(zF_w(z)) &= \sum_{n=0}^{\infty} n P_{n-1}(w) z^n \,. \end{aligned}$$

Based on the recursion formula (2.2.67) we can deduce

$$\begin{aligned} \frac{dF_w}{dz}(z) - (2w\frac{d}{dz}(zF_w(z)) - wF_w(z)) + z\frac{d}{dz}(zF_w(z)) &= 0 \,, \\ \frac{dF_w}{dz}(z) &= \frac{w-z}{z^2 - 2wz + 1} F_w(z) \,. \end{aligned} \tag{2.2.71}$$

By the substitution $z \leftarrow t$: ODE for $t \mapsto F_w(t)$.

Corresponding initial value problem with $F_w(0) = P_0(w) = 1$ has unique solution

$$F_w(z) = \left(z^2 - 2wz + 1\right)^{-1/2} .$$

Hereby, we substituted again $t \leftarrow z$ in the sense of analytic continuation. Therefore, there holds for fixed $w \in \mathbb{C}$ and $|z|$ sufficiently small

$$\left(z^2 - 2wz + 1\right)^{-1/2} = \sum_{n=0}^{\infty} P_n(w) z^n . \tag{2.2.72}$$

Generating function of the Legendre polynomials

Factorization of $F_w(z)$: with

$$w := \tfrac{1}{2}(\zeta + \zeta^{-1}) , \quad \zeta \in \mathbb{C} \setminus \{0\} \quad \Rightarrow \quad z^2 - 2wz + 1 = (1 - z\zeta)(1 - z/\zeta) . \tag{2.2.73}$$

From the Taylor series for $(1-z)^{-1/2}$: ($\rightarrow$ analysis), for $|z| < 1$

$$(1-z)^{-1/2} = \sum_{n=0}^{\infty} a_n z^n , \quad a_n = \frac{(2n)!}{(n!)^2 2^{2n}} = \frac{1 \cdot 2 \cdot \cdots \cdot 2n}{(2 \cdot 4 \cdot \cdots \cdot 2n)^2} > 0 .$$

From the multiplication theorem for power series with transformation (2.2.72), for $|z| < \min\{|\zeta|, |\zeta|^{-1}\}$

$$\left(z^2 - 2wz + 1\right)^{-1/2} = (1 - z\zeta)^{-1/2}(1 - z/\zeta)^{-1/2} = \sum_{n=0}^{\infty} \left(\sum_{j=0}^{n} a_j a_{n-j} \zeta^{n-2j} \right) z^n .$$

Lemma 2.2.74 (Upper bound for Legendre polynomials)**.**

There holds $|P_n(t)| \leq 1$ *for all* $-1 \leq t \leq 1$.

Proof. We use the representation of the nth Legendre polynomial from the series expansion of the generating function: with the Joukowski transformation $T(\zeta) = \frac{1}{2}(\zeta + \zeta^{-1})$ there holds

$$P_n(T(\zeta)) = \sum_{j=0}^{n} a_j a_{n-j} \zeta^{n-2j} \ , \quad |\zeta| \geq 1 \ . \tag{2.2.75}$$

Note: $\frac{1}{2}(e^{i\varphi} + e^{-i\varphi}) = \cos\varphi =: t$

$$\Rightarrow \quad \text{For} \quad T(\zeta) := \frac{1}{2}(\zeta + \zeta^{-1}) \quad \text{there holds} \quad T(\{|z| = 1\}) = [-1, 1] \ .$$

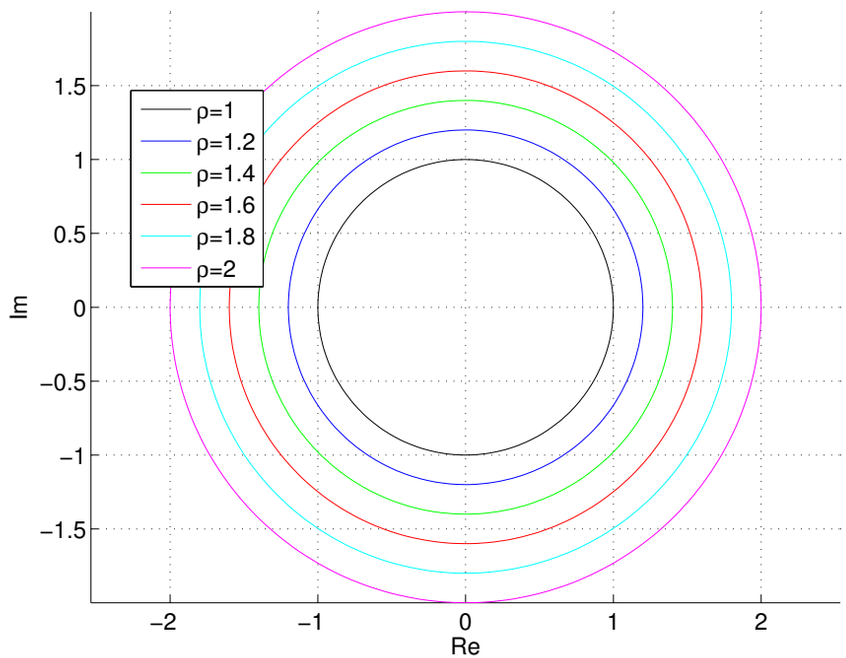
$$\Rightarrow \quad |P_n(t)| \overset{(2.2.74)}{=} \left| \sum_{j=0}^{n} a_j a_{n-j} \exp(i(n-2j)\varphi) \right| \overset{a_j > 0}{\leq} \sum_{j=0}^{n} a_j a_{n-j} = P_n(1) = 1 \ ,$$

for a $\varphi \in [0, 2\pi]$ such that $T(\exp(i\varphi)) = t \in [-1, 1]$. □

We now consider the effect of the Joukowski-Transformation $T(z) = \frac{1}{2}(z+z^{-1})$ on the circles around 0 more closely

$$T(\rho e^{i\varphi}) = \tfrac{1}{2}\left(\rho e^{i\varphi} + \rho^{-1}e^{-i\varphi}\right) = \underbrace{\frac{1}{2}(\rho+\rho^{-1})}_{\text{semi-major axis}}\ \cos(\varphi) + i\cdot\underbrace{\frac{1}{2}(\rho-\rho^{-1})}_{\text{semi-minor axis}}\ \sin(\varphi)\ .$$

➢ T maps a circle with radius $\rho > 1$ to an ellipse with foci $\{-1,1\}$, semi-minor axis $\frac{1}{2}(\rho-\rho^{-1})$, and semi-major axis $\frac{1}{2}(\rho+\rho^{-1})$.



Circles $\{z \in \mathbb{C} : \ |z| = \rho\}$

Transformation T

$\xrightarrow{z \to \frac{1}{2}(z+1/z)}$

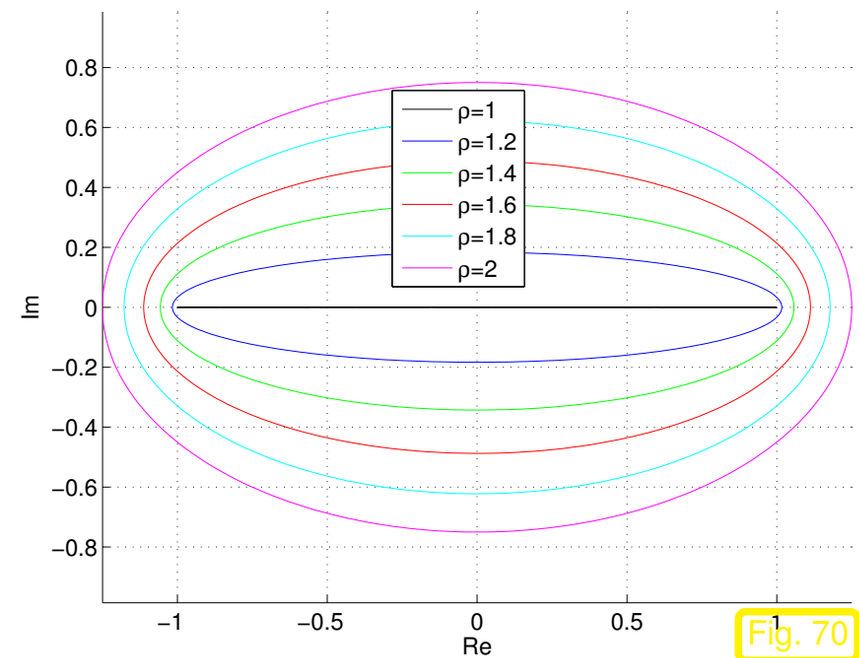
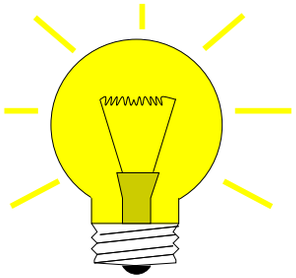


Fig. 70

Ellipses E_ρ according to (2.2.75)

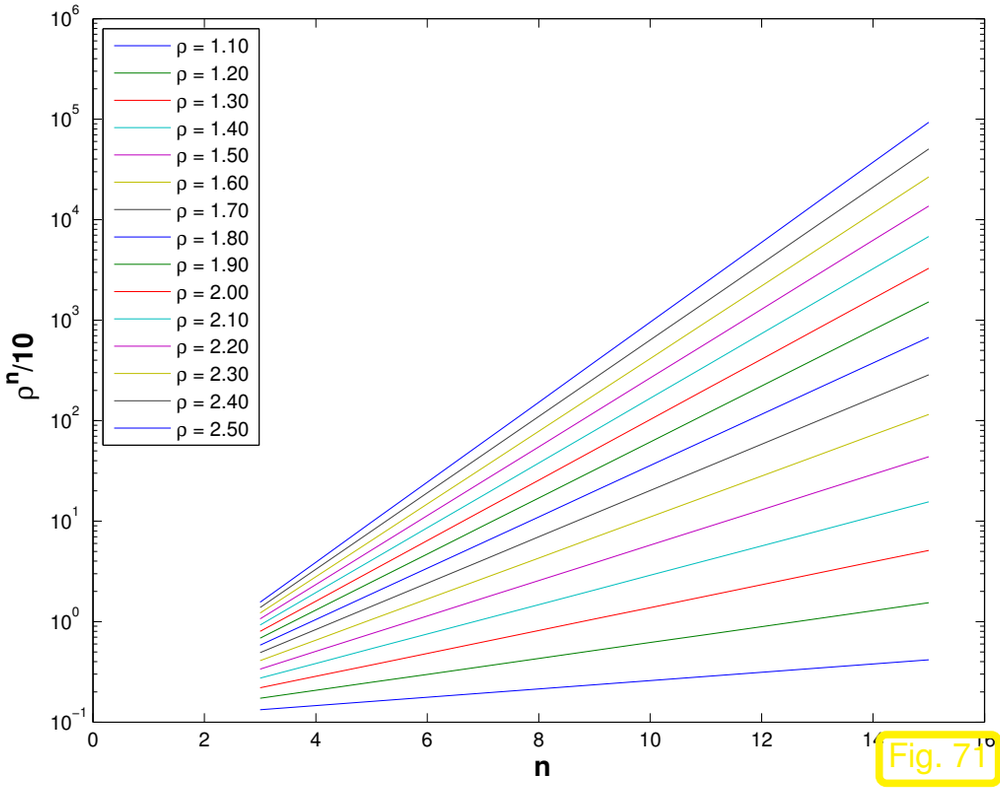
Figures 67-69: level lines of $|P_n(z)|$ are approximately ellipses with foci $\{-1,1\}$.



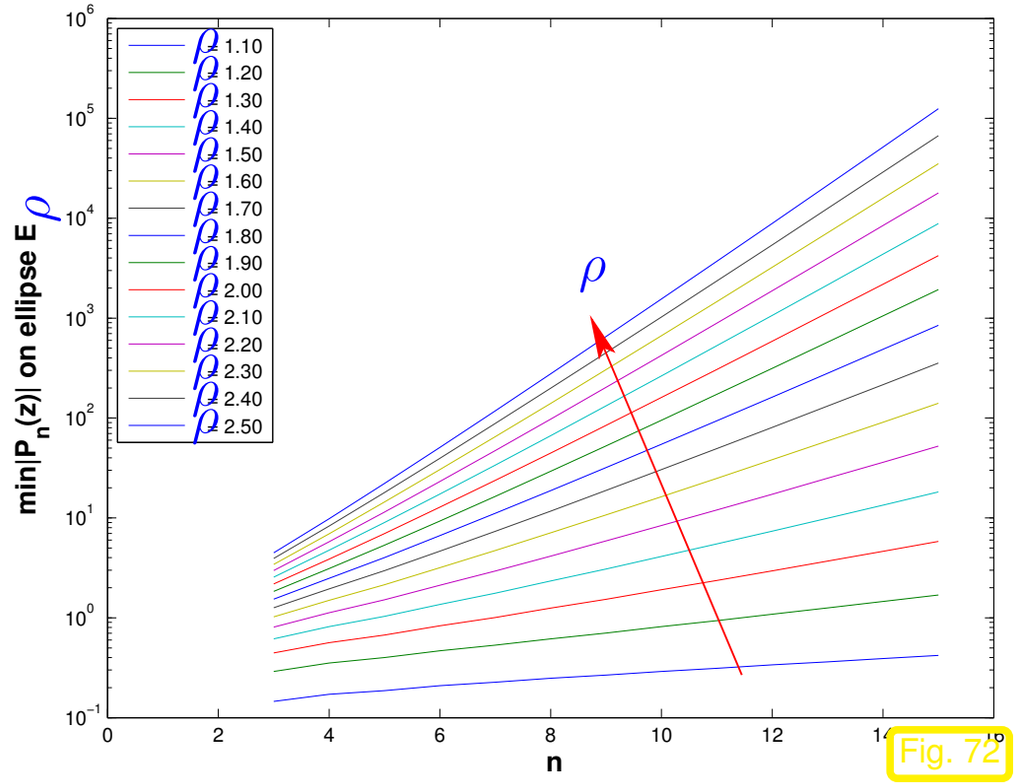
Idea: Use elliptic integration path, compare Fig. 70

$$E_\rho := T(\{z \in \mathbb{C},\ |z| = \rho\})\ , \quad \rho > 1\ , \tag{2.2.76}$$

with Joukowski transformation $T(z) := \frac{1}{2}(z + 1/z)$. (2.2.77)



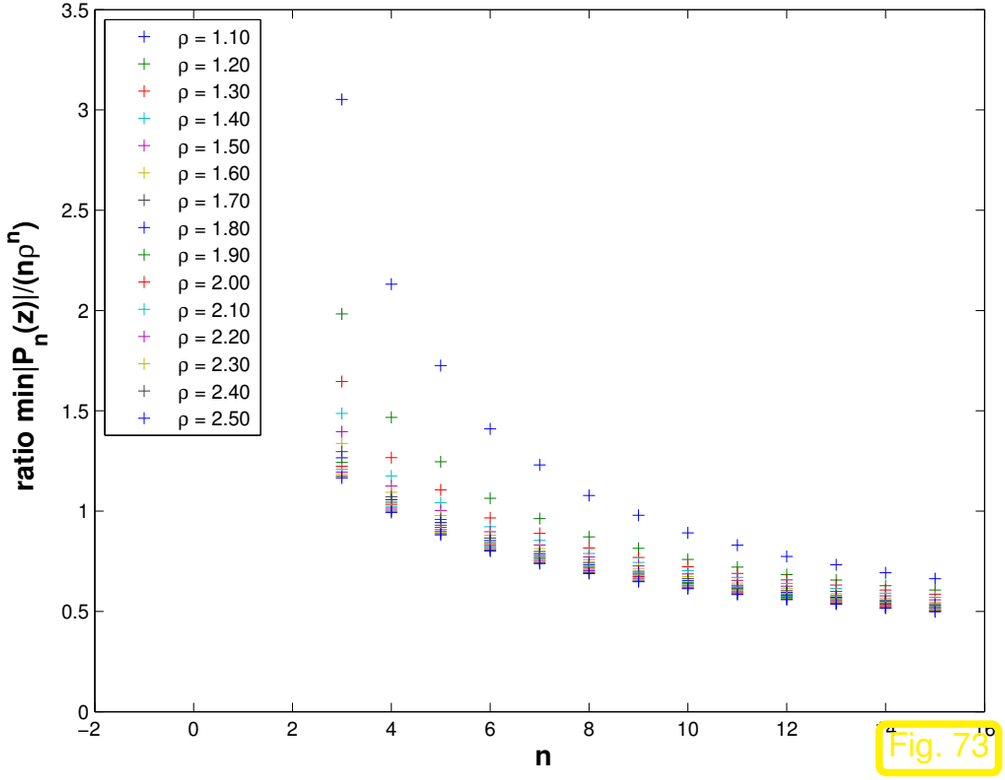
$0.1 \cdot \rho^n$ as function of n for different $\rho > 1$



Approximation of $\min_{z \in E_\rho} |P_n(z)|$

P. Grohs

rev 63606, February 17, 2014



Conjecture:

for large n:

$$\min_{z \in E_\rho} |P_n(z)| \sim n\rho^n \,. \tag{2.2.78}$$

P. Grohs

rev 63606, February 17, 2014

To the best knowledge of the author, there is no proof of (2.2.77). In [8, Sect. 12.4], [7] the following weaker claim is shown

$$\forall \epsilon > 0\colon \quad \exists N = N(\epsilon)\colon \min_{z \in E_\rho} |P_n(z)| \geq (\rho - \epsilon)^n \quad \forall n > N(\epsilon).$$

Here only *heuristically*:

Reminder: Formula of Cauchy-Hadamard for the radius of convergence of a power series [36, Sect. 4.1.3]

$$R := \frac{1}{\limsup\limits_{n\to\infty} |a_n|^{1/n}} \quad \Rightarrow \quad \sum_{n=0}^{\infty} a_n z^n \quad \text{converges for } |z| < R\,. \tag{2.2.79}$$

Idea: (2.2.78) implies asymptotic lower bound for $|a_n|$, if R is known !
(Prerequisite is $\limsup\limits_{n\to\infty} |a_n|^{1/n} = \lim\limits_{n\to\infty} |a_n|^{1/n}$; this is for example the case for monotonicity of $(a_n)_n$.)

Based on the expansion theorem of Cauchy one can deduce, compare [36, Sect. 8.1.5]:

Lemma 2.2.80 (Radius of convergence of power series expansions)**.**
Let $D \subset \mathbb{C}$ be open, $f : D \mapsto \mathbb{C}$ holomorphic and $0 \in D$. Then, the Taylor series of f around 0 has the radius of convergence $R = \operatorname{dist}(0, \partial D)$.

▶ To examine using Lemma 2.2.79: radius of convergence of the Taylor series around 0 of the generating function (2.2.71), i.e., $f(z) = (z^2 - 2wz + 1)^{-1/2}$, of the Legendre polynomials from (2.2.71).

We are interested in: $\min_{z\in E_\rho} |P_n(z)|$ ➥ Wanted: radius of convergence for $w \in E_\rho$

$$\text{Lemma 2.2.79} \quad \Rightarrow \quad R = \min\{\mathrm{dist}(0,\zeta), \mathrm{dist}(0,\zeta^{-1})\} \ ,$$

as ζ, ζ^{-1} are the two roots of $z \mapsto z^2 - 2wz + 1$ for $w = \frac{1}{2}(\zeta + \zeta^{-1})$, cmp. (2.2.72).

$$\blacktriangleright \qquad \zeta = \rho \exp(i\varphi) \ , \quad \rho > 1 \quad \Rightarrow \quad R = \rho^{-1} \ .$$

Assumption. Existence of limit: $\limsup_{n\to\infty} |P_n(w)|^{1/n} = \lim_{n\to\infty} |P_n(w)|^{1/n}$

Therefore from (2.2.71):

$$\forall \epsilon > 0\text{:} \quad \exists N = N(\epsilon) \in \mathbb{N}\text{:} \quad \left(|P_n(w)|^{1/n} > \rho - \epsilon \quad \Leftrightarrow \quad |P_n(w)| > (\rho - \epsilon)^n \right) \quad \forall n > N(\epsilon) \ .$$

Theorem 2.2.81 (Error estimate for interpolation in Gauss nodes)**.**
Let there be an analytic continuation of $f : [-1,1] \mapsto \mathbb{C}$ *to* $D \subset \mathbb{C}$ *and let* $E_\rho \subset D$ *for a* $\rho > 1$*.*
Under the <u>*assumption*</u> (2.2.77) *there exists* $N = N(\rho) \in \mathbb{N}$ *and* $C = C(\rho) > 0$*, such that*

$$\max_{-1\le t\le 1} \left| f(t) - \sum_{j=1}^{s} f(\tau_j) L_j(t) \right| \le C \frac{\rho^{-s}}{s} \max_{z\in E_\rho} |f(z)| \frac{\mathrm{length}(E_\rho)}{2\pi} \quad \forall s > N(\rho) \ .$$

Content of Theorem 2.2.80: w.r.t. the maximum norm exponential convergence ($\to$ Def. 1.4.5) of the interpolation polynomial in the Gauss nodes of a function that is analytic in a "suitable" neighborhood of $[-1,1]$.

Example 2.2.82 (Error for polynomial interpolation in Gauss nodes)*.* $\to$ Continuation Ex. 2.2.58

$$f(z) = \frac{1}{1+\exp(-\frac{1}{2}\lambda z)} \quad \Rightarrow \quad f \text{ analytic in } E_\rho \text{ for } \rho < \frac{\pi}{\lambda} + \sqrt{(\frac{\pi}{\lambda})^2 + 1} \ .$$

Pole of f:

$$\pm(2k+1)\frac{2\pi i}{\lambda}\,, \quad k \in \mathbb{Z}\,.$$

$$\Rightarrow \quad \rho - \rho^{-1} < \frac{2\pi}{\lambda}\,.$$

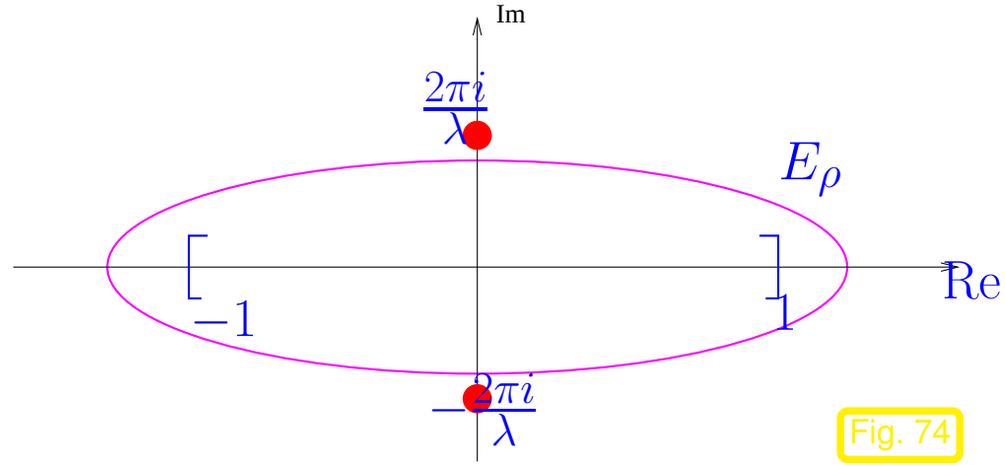


Fig. 74

This example demonstrates the general strategy for finding admissible ellipses of analyticity for "simple" functions: one determines the poles $\mathbb{C}$ of the function that is examined and thereby the domain, on which the function is holomorphic.

$\diamond$

In rare cases, for example for the logistic differential equation from Ex. 1.2.1, we can use the analytically given solution of the IVP to determine the domain of analyticity, in order to confirm the exponential convergence of the global Gauss collocation method.

In general, this information is missing. Nevertheless, statements concerning the domain of analyticity of solutions of IVPs for differential equations in $\mathbb{C}$ with locally holomorphic right hand side are possible:

Theorem 2.2.83 (Existence and uniqueness for differential eqns in $\mathbb{C}$). $\rightarrow$ *[38, Kap. I, §8]*
Let $f : D \subset \mathbb{C} \mapsto \mathbb{C}$ be holomorphic in a neighborhood $B_\rho(z_0) := \{z \in \mathbb{C} : |z - z_0| < \rho\} \subset D$ of $z_0 \in D$ and let $|f(z)| \leq M$ for all $z \in B_\rho(z_0)$. Then there exists exactly one solution y that is holomorphic on $B_{\rho/M}(0)$ and solves the initial value problem

$$y'(z) = f(y(z)) \quad \forall z \in B_{\rho/M}(0) \quad , \quad y(0) = z_0 \, .$$

Notation: $' \hat{=}$ complex differentiation

If $f(z) \in \mathbb{R}$ for $z \in \mathbb{R}$ and $y_0 \in \mathbb{R}$, then the locally holomorphic solution of the complex IVP with real arguments postulated by the theorem coincides with the solution according to theorem 1.3.4.

[Beitrag von J. Mitrovic, FS 2011]

Examples of ODEs with holomorphic right hand side: Lotka-Volterra ODE, pendulum equation, logistic ODE

2.3 Runge-Kutta methods

Disadvantage of the single step collocation methods: All (with the exception of the explicit Euler method) are *implicit* ($\rightarrow$ Def. 2.1.5)

Are there explicit single step methods of higher order **?** If so, how do we find them **?**

2.3.1 Construction

IVP: $$\begin{aligned}\dot{\mathbf{y}}(t) &= \mathbf{f}(t,\mathbf{y}(t))\,,\\ \mathbf{y}(t_0) &= \mathbf{y}_0\end{aligned} \quad\Rightarrow\quad \mathbf{y}(t_1) = \mathbf{y}_0 + \int_{t_0}^{t_1} \mathbf{f}(\tau,\mathbf{y}(\tau))\,\mathrm{d}\tau$$

Approximation by quadrature formula (on $[0,1]$) with s nodes $c_1,\dots,c_s$:

$$\mathbf{y}(t_1) \approx \mathbf{y}_1(=\mathbf{y}_h(t_1)) = \mathbf{y}_0 + h\sum_{i=1}^{s} b_i\mathbf{f}(t_0+c_ih, \boxed{\mathbf{y}(t_0+c_ih)})\,, \quad h := t_1 - t_0\,.$$

How do we get these values **?** $\triangleright$ Bootstrapping

Example 2.3.1 (Construction of simple Runge-Kutta methods).

Quadrature formula $\to$ trapezoidal rule: on interval $[a,b]$

$$Q(f) = \tfrac{1}{2}(b-a)(f(a)+f(b)) \quad\leftrightarrow\quad s=2\colon\ c_1=0, c_2=1\,,\quad b_1=b_2=\frac{1}{2}\,, \tag{2.3.2}$$

and $\mathbf{y}_h(T)$ from explicit Euler step (1.4.2)

$$\mathbf{k}_1 = \mathbf{f}(t_0,\mathbf{y}_0)\,,\quad \mathbf{k}_2 = \mathbf{f}(t_0+h,\mathbf{y}_0+h\mathbf{k}_1)\,,\quad \mathbf{y}_1 = \mathbf{y}_0 + \tfrac{h}{2}(\mathbf{k}_1+\mathbf{k}_2)\,. \tag{2.3.3}$$

(2.3.3) **=** explicit trapezoidal rule

Quadrature formula $\rightarrow$ simplest Gauss quadrature formula (midpoint rule) **&** $\mathbf{y}_h(\frac{1}{2}(t_1+t_0))$ from explicit Euler step (1.4.2)

$$\mathbf{k}_1 = \mathbf{f}(t_0, \mathbf{y}_0) \;, \quad \mathbf{k}_2 = \mathbf{f}(t_0 + \tfrac{h}{2}, \mathbf{y}_0 + \tfrac{h}{2}\mathbf{k}_1) \;, \quad \mathbf{y}_1 = \mathbf{y}_0 + h\mathbf{k}_2 \;. \tag{2.3.4}$$

(2.3.4) **=** explicit midpoint rule

$\diamond$

Discrete evolutions of the form (2.2.3)

Definition 2.3.5 (Runge-Kutta methods)**.**
For $b_i, a_{ij} \in \mathbb{R}$, $c_i := \sum_{j=1}^{s} a_{ij}$, $i,j = 1,\dots,s$, $s \in \mathbb{N}$, *the formula*

$$\mathbf{k}_i := \mathbf{f}(t_0 + c_i h, \mathbf{y}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j) \;, \quad i = 1,\dots,s \quad , \quad \mathbf{\Psi}^{t_0,t_0+h}\mathbf{y}_0 := \mathbf{y}_0 + h\sum_{i=1}^{s} b_i\mathbf{k}_i \;,$$

defines a s*-stage (single step) Runge-Kutta method (RK-SSM) for IVP* (1.1.13) *with increments* $\mathbf{k}_i \in \mathbb{R}^d$.

➢ Generalization of the collocation method → Sect. 2.2
(but no concrete rules for construction !)

If $a_{ij} = 0$ for $i \leq j$ ▶ explicit Runge-Kutta method → Def. 2.1.5

Short notation for Runge-Kutta method:

Butcher tableau ▷

$$\begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^T \end{array} := \begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array} . \tag{2.3.6}$$

$\mathfrak{A}$ strictly lower triangular matrix ➤ explicit Runge-Kutta method
$\mathfrak{A}$ lower triangular matrix ➤ diagonally implicit Runge-Kutta method (DIRK)

P. Grohs

rev 63606, February 17, 2014

Remark 2.3.7 (Stage form of increment equations).

for s-stage Runge-Kutta method (RK-SSM), → Def. 2.3.5, define (assumption: unique solvability of increment equations)

Stages
$$\mathbf{g}_i = \mathbf{y}_0 + h \sum_{j=1}^{s} a_{ij} \mathbf{k}_j \ , \quad i = 1, \ldots, s \quad \Rightarrow \quad \mathbf{k}_i = \mathbf{f}(t_0 + c_i h, \mathbf{g}_i) \ . \tag{2.3.8}$$

Equations for stages

$$\mathbf{g}_i = \mathbf{y}_0 + h \sum_{j=1}^{s} a_{ij} \mathbf{f}(t_0 + c_j h, \mathbf{g}_j) \quad , \quad i = 1, \dots, s \; . \tag{2.3.9}$$

Interpretation: Runge-Kutta method $\leftrightarrow$ approximation of the solution curve by polygonal lines

$\rightarrow$ Sect. 1.4

- Number $b_i \neq 0 \mathrel{\hat{=}}$ number of legs in the polygonal line
- b_i, $i = 1, \dots, s-1 \mathrel{\hat{=}}$ relative length of the ith interval part
- $\mathbf{k}_i \mathrel{\hat{=}}$ "slope" of the ith interval part
- $c_i \mathrel{\hat{=}}$ relative point in time (in $[t_k, t_{k+1}]$) for evaluation of the slope of the ith part

Example 2.3.10 (Explicit Runge-Kutta polygonal line approximation for Ricatti differential equation). $\rightarrow$ Ex 1.1.3

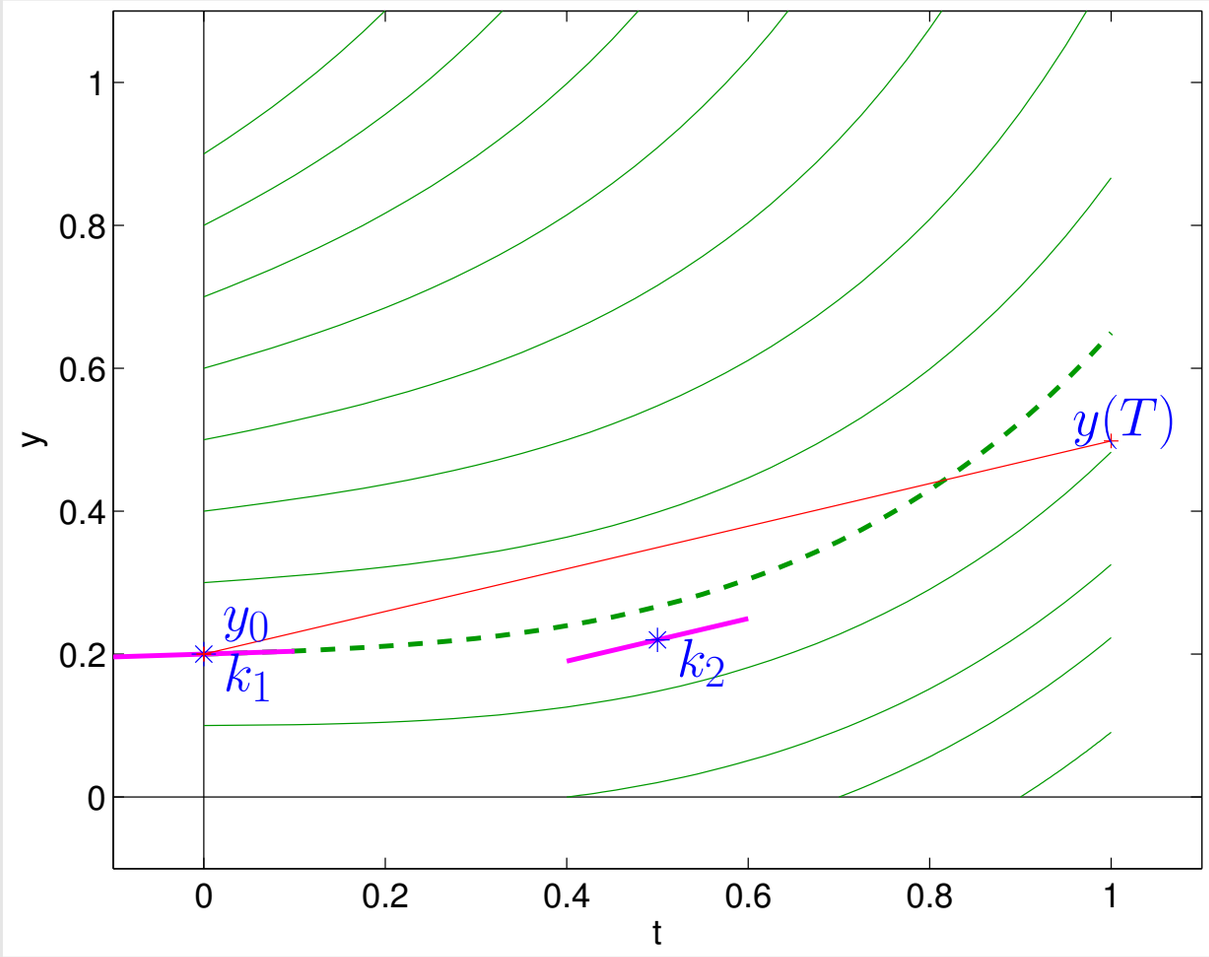
initial value problem: $\dot{y} = t^2 + y^2$, $y(0) = 0.2$.

Geometric interpretation of explicit RK-SSM as polygonal line method $\rightarrow$ Generalization of the explicit Euler method, compare Sect. 1.4.1, Fig. **??**.

explicit midpoint rule:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array}$$

green	solution curve
magenta	slope of interval part $\mathbf{k}_i$
$*$	point f-evaluation
red:	polygonal line



P. Grohs

rev 63606, February 17, 2014

explicit trapezoidal rule

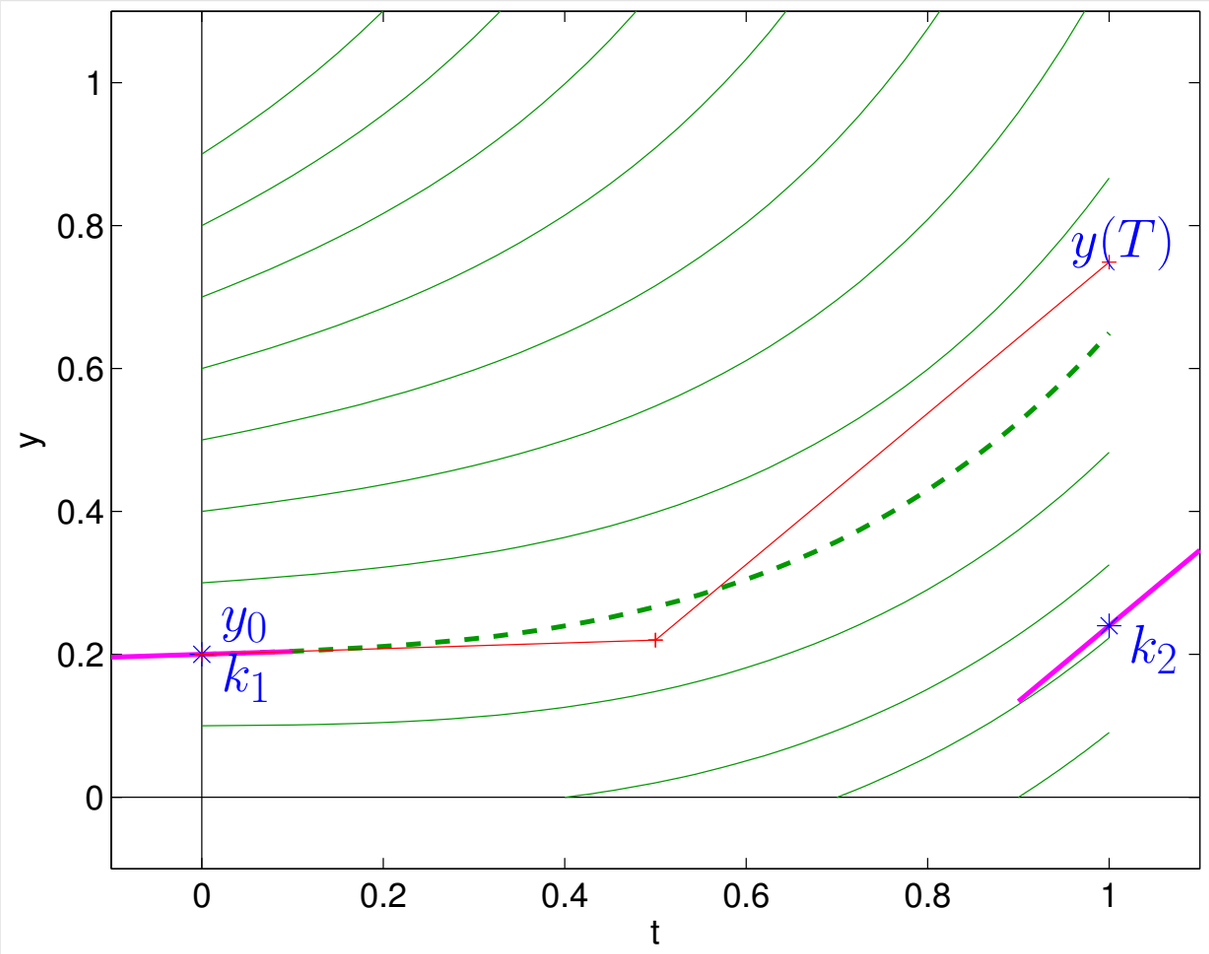
$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

green: solution curve

magenta: slope of interval part $\mathbf{k}_i$

$*$: point f-evaluation

red: polygonal line



P. Grohs

rev 63606, February 17, 2014

Classic Runge-Kutta method (RK4)

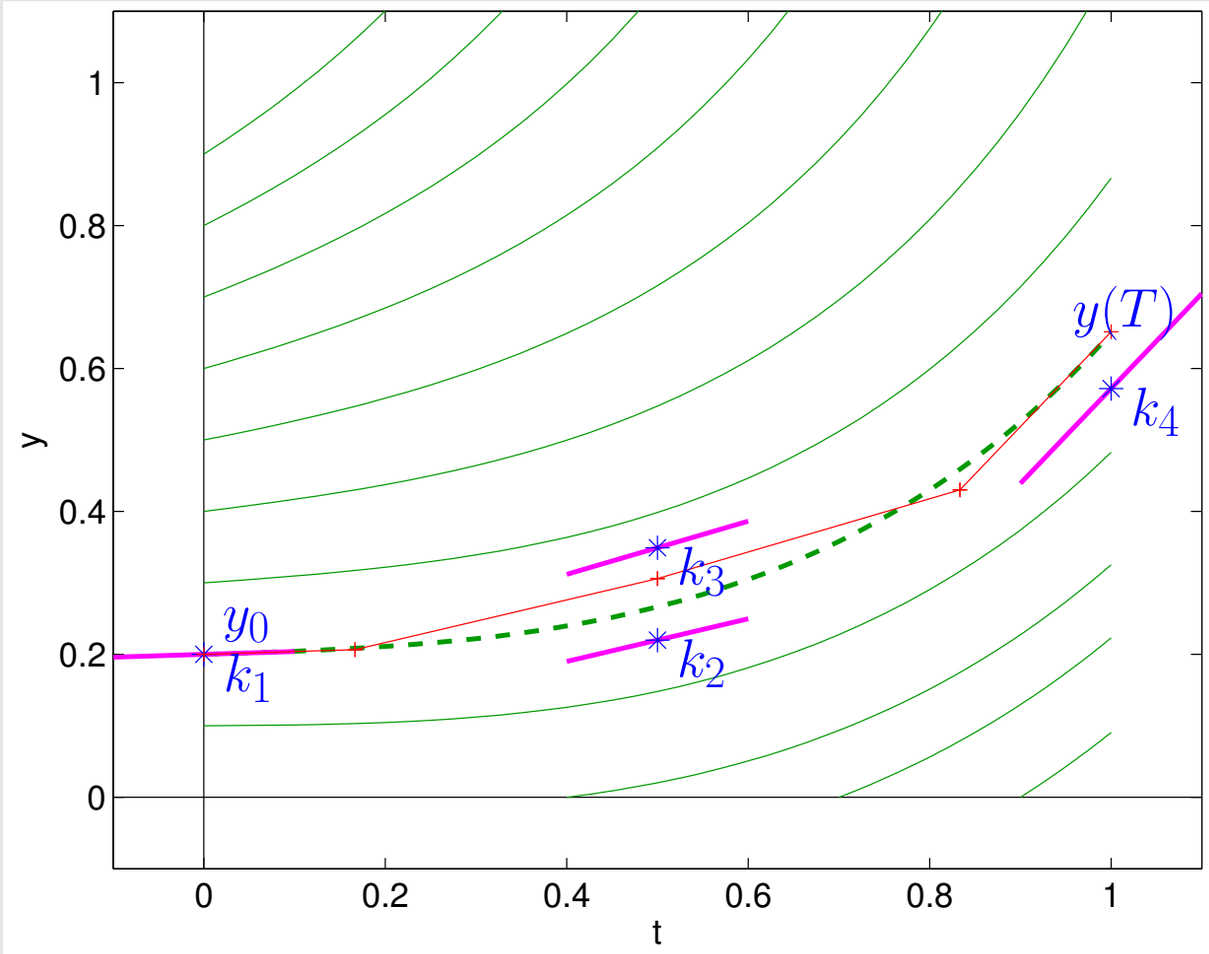
$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array} \qquad (2.3.11)$$

green solution curve

magenta slope of interval part $\mathbf{k}_i$

$*$ point f-evaluation

red: polygonal line



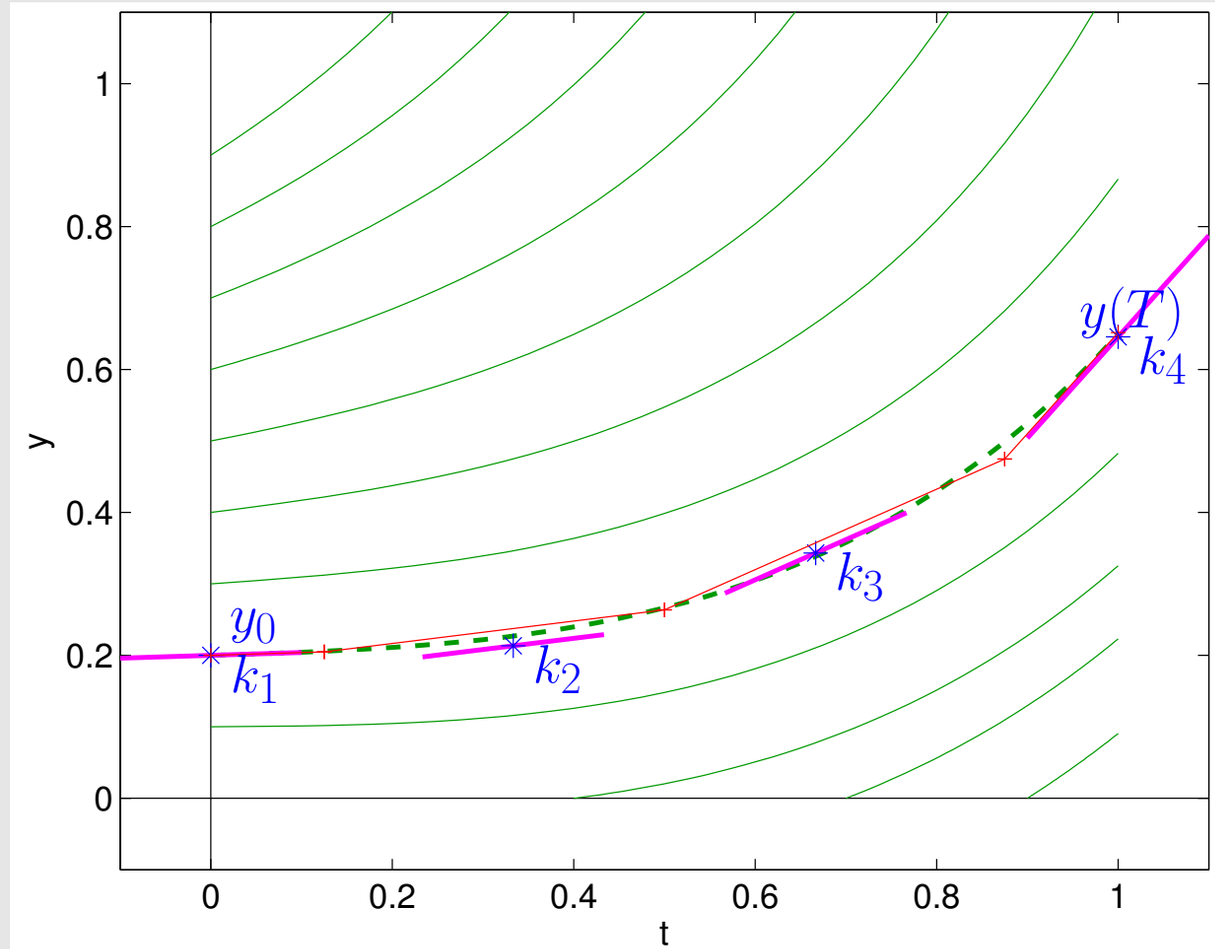
P. Grohs

rev 63606, February 17, 2014

Kuttas 3/8-rule

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{2}{3} & -\frac{1}{3} & 1 & 0 & 0 \\ 1 & 1 & -1 & 1 & 0 \\ \hline & \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{array} \tag{2.3.12}$$

green	solution curve
magenta	slope of interval part $\mathbf{k}_i$
$*$	point f-evaluation
red:	polygonal line



◇

Remark 2.3.13 (Affine covariance of the Runge-Kutta method).

How does a Runge-Kutta SSM react to a change of basis in the state space (→ Sect. 1.3.2, (1.3.12))?

For $\mathbf{S} \in \mathbb{R}^{d,d}$ regular, $\widehat{\mathbf{y}} := \mathbf{S}^{-1}\mathbf{y}$ $\quad \mathbf{\Psi}$, $\widehat{\mathbf{\Psi}}$ from RK methods (→ Def. 2.3.5)

$$\begin{aligned} \mathbf{\Psi}_h^{s,t} &= \text{ discrete evolution to } \quad \dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y}) \ , \\ \widehat{\mathbf{\Psi}}_h^{s,t} &= \text{ discrete evolution to } \quad \dot{\widehat{\mathbf{y}}} = \widehat{f}(t,\widehat{\mathbf{y}}) \quad \to (1.3.12) \end{aligned} \quad \blacktriangleright \quad \boxed{\mathbf{S}\widehat{\mathbf{\Psi}}_h^{s,t}\mathbf{S}^{-1}\mathbf{y} = \mathbf{\Psi}_h^{s,t}\mathbf{y}} \ . \tag{2.3.14}$$

One needs to show that the increments $\mathbf{k}_i$ and the transformed increments $\mathbf{S}\widehat{\mathbf{k}}_i\mathbf{S}^{-1}$ of a Runge-Kutta SSM applied to the ODE $\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y})$, and $\dot{\widehat{\mathbf{y}}} = \widehat{\mathbf{f}}(t,\widehat{\mathbf{y}}) = \mathbf{S}^{-1}\mathbf{f}(t,\mathbf{S}\widehat{\mathbf{y}})$, respectively, fulfill the same equations. Due to the unique solvability for sufficiently small $h > 0$ this implies $\mathbf{k}_i = \mathbf{S}\widehat{\mathbf{k}}_i\mathbf{S}^{-1}$, $i = 1,\dots,s$.

P. Grohs

rev 63606, February 17, 2014

The above statement can be expressed by a *commutative diagram*

$$\begin{tikzcd} \dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y}) \arrow[r, "\text{Change of basis } \widehat{\mathbf{y}}:=\mathbf{S}^{-1}\mathbf{y}"] \arrow[d, "\text{RK SSM}"'] & \dot{\widehat{\mathbf{y}}} = \widehat{\mathbf{f}}(t,\widehat{\mathbf{y}}) \arrow[d, "\text{RK SSM}"] \\ (\mathbf{y}_k)_k \arrow[r, "\text{Change of basis } \widehat{\mathbf{y}}_k:=\mathbf{S}^{-1}\mathbf{y}_k"] & (\widehat{\mathbf{y}}_k)_k \ . \end{tikzcd}$$

☞ Affine covariance expresses the *preservation* of a simple algebraic *structure* of the solution set of an IVP.

Remark 2.3.15 (Invariance w.r.t. autonomization of a Runge-Kutta method).

Another transformation of an ODE: autonomization $\rightarrow$ Rem. 1.1.7

Also here, the question arises when "RK-SSM commute with autonomization", cmp. Rem. 2.3.13.

Autonomization: $\begin{array}{c}\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y})\,,\\ \mathbf{y}(t_0) = \mathbf{y}_0\end{array} \quad \Rightarrow \quad \dot{\mathbf{z}} := \dot{\begin{pmatrix}\mathbf{y}\\ s\end{pmatrix}} = \begin{pmatrix}\mathbf{f}(s,\mathbf{y})\\ 1\end{pmatrix} =: \mathbf{g}\begin{pmatrix}\mathbf{y}\\ s\end{pmatrix}, \quad \begin{pmatrix}\mathbf{y}(0)\\ s(0)\end{pmatrix} = \begin{pmatrix}\mathbf{y}_0\\ t_0\end{pmatrix}.$

$\blacktriangledown \qquad\qquad\qquad\qquad \blacktriangledown$

Evolutions: $\mathbf{\Phi}^{t,t+h} \quad \leftrightarrow \quad \widehat{\mathbf{\Phi}}^{h}$,

Discrete evl.: $\mathbf{\Psi}_h^{t,t+h} \quad \leftrightarrow \quad \widehat{\mathbf{\Psi}}_h^{h}$.

Wish:
$$\begin{pmatrix}\mathbf{\Phi}^{t,t+h}\mathbf{y}\\ t+h\end{pmatrix} = \widehat{\mathbf{\Phi}}^{h}\begin{pmatrix}\mathbf{y}\\ t\end{pmatrix} \quad \blacktriangleright \quad \begin{pmatrix}\mathbf{\Psi}_h^{t,t+h}\mathbf{y}\\ t+h\end{pmatrix} = \widehat{\mathbf{\Psi}}_h^{h}\begin{pmatrix}\mathbf{y}\\ t\end{pmatrix}. \tag{2.3.16}$$

(2.3.16) can again be expressed by a *commutative diagram*:

$$\begin{tikzcd}
\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y}) \arrow[r, "\text{Autonomization}"] \arrow[d, "\text{RK SSM}"'] & \dot{\mathbf{z}} = \mathbf{g}(\mathbf{z}) \arrow[d, "\text{RK SSM}"] \\
(\mathbf{y}_k)_k \arrow[r, "\widehat{\mathbf{y}}_k := \binom{\mathbf{y}_k}{t_k}"] & (\widehat{\mathbf{y}}_k)_k\,.
\end{tikzcd}$$

Now we want to deduce conditions for the coefficients a_{ij} and b_i of the RK SSM ($\rightarrow$ Def. 2.3.5) such that there holds (2.3.16).

$$\widehat{\mathbf{\Psi}}_h^h \begin{pmatrix} \mathbf{y} \\ t \end{pmatrix} = \begin{pmatrix} \mathbf{y} + h\sum_{i=1}^s b_i \widehat{\mathbf{k}}_i \\ t + h\sum_{i=1}^s b_i \widehat{\kappa}_i \end{pmatrix} , \quad \begin{pmatrix} \widehat{\mathbf{k}}_i \\ \widehat{\kappa}_i \end{pmatrix} = \begin{pmatrix} \mathbf{f}(t + h\sum_{j=1}^s a_{ij}\widehat{\kappa}_j, \mathbf{y} + h\sum_{j=1}^s a_{ij}\widehat{\mathbf{k}}_j) \\ 1 \end{pmatrix} .$$

$$\boxed{c_i = \sum_{j=1}^s a_{ij}} \quad \& \quad \boxed{\sum_{i=1}^s b_i = 1} \quad \blacktriangleright \quad \widehat{\mathbf{k}}_i = \mathbf{k}_i \ . \tag{2.3.17}$$

= Sufficient + necessary conditions for invariance w.r.t. autonomization of a RK method

Therefore $c_i = \sum_{j=1}^s a_{ij}$ in Def. 2.3.5 **!**

▷ Analysis of RK methods that are invariant w.r.t. autonomization can be constrained to autonomous problems.

△

Remark 2.3.18 ("'Dense output"). [22, Sect. II.5]

(Single step) Runge-Kutta methods ($\rightarrow$ Def. 2.3.5) produce grid functions $\mathcal{G} \mapsto \mathbb{R}^d$ as approximation of $t \mapsto \mathbf{y}(t)$ at discrete points in time.

What happens if one needs approximations of $\mathbf{y}(t)$ at other points in time/everywhere on $[0,T]$?

Goal:
- Piecewise polynomial definition of $t \mapsto \mathbf{y}_h(t)$
- Interpolation property $\mathbf{y}_h(t_k) = \mathbf{y}_k$, $k = 0, \dots, N$
- $\mathbf{y}_{h|[t_k,t_{k+1}}$ computable from $\mathbf{y}_k$, $\mathbf{y}_{k+1}$ and increments in k^{th} step

$$\blacktriangleright \quad \mathbf{y}_h(t_k + \xi h_k) = p_0(\xi)\mathbf{y}_k + p_1(\xi)\mathbf{y}_{k+1} + \sum_{i=1}^{s} q(\xi)\mathbf{k}_i \ , \quad 0 \le \xi \le 1 \ ,$$

with polynomials $p_0, p_1, q_i : \mathbb{R} \mapsto \mathbb{R}$.

Wish: for RK SSM of the order p $\quad\succ\quad \max\limits_{0 \le t \le T} \|\mathbf{y}(t) - \mathbf{y}_h(t)\| = O(h^p)$

$\triangle$

Remark 2.3.19 (Solution of the increment equations). $\rightarrow$ [12, Sect. 6.2.2]

Increment equations for implicit RK SSM ($\rightarrow$ Def. 2.3.5) = (in general *nonlinear*) system of equations with $s \cdot d$ unknowns

In the autonomous case (cmp. proof of Lemma 2.2.7)

$$\mathbf{k}_i := \mathbf{f}(\mathbf{y}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j) \quad \overset{\mathbf{k}_i=\mathbf{f}(\mathbf{y}_0+\mathbf{g}_i)}{\Longleftrightarrow} \quad \mathbf{g}_i = h\sum_{j=1}^{s} a_{ij}\mathbf{f}(\mathbf{y}_0 + \mathbf{g}_j)\,, \quad i = 1,\dots,s\,. \tag{2.3.20}$$

The quantities $\mathbf{g}_i + \mathbf{y}_0$ are called stages of the Runge-Kutta method, compare Rem. 2.3.7. Therefore, one calls the formulation of the increment equations by means of $\mathbf{g}_i$ also their stage formulation as in (2.3.20).

➢ iterative solution with *simplified* Newton method ("frozen" Jacobian matrix)

! Efficiency: Minimize number of $\mathbf{f}$, $D\mathbf{f}$-evaluations

With $\mathfrak{g} = (\mathbf{g}_1,\dots,\mathbf{g}_s)^T \in \mathbb{R}^{s\cdot d}$ define

$$F(\mathfrak{g}) := \mathfrak{g} - \begin{pmatrix} h\sum_{j=1}^{s} a_{1j}\mathbf{f}(\mathbf{y}_0+\mathbf{g}_j) \\ \vdots \\ h\sum_{j=1}^{s} a_{sj}\mathbf{f}(\mathbf{y}_0+\mathbf{g}_j) \end{pmatrix} \quad \Rightarrow \quad \{(2.3.20) \quad \Leftrightarrow \quad F(\mathfrak{g}) = 0\}\,. \tag{2.3.21}$$

h "'small" ➢ Natural initial approximation for simplified Newton iteration: $\mathfrak{g}^{(0)} = 0$

$$\blacktriangleright \quad DF(\mathfrak{g}^{(0)}) = \begin{pmatrix} \mathbf{I} - ha_{11}D\mathbf{f}(\mathbf{y}_0) & -ha_{12}D\mathbf{f}(\mathbf{y}_0) & \cdots & -ha_{1s}D\mathbf{f}(\mathbf{y}_0) \\ -ha_{21}D\mathbf{f}(\mathbf{y}_0) & \mathbf{I} - ha_{22}D\mathbf{f}(\mathbf{y}_0) & & \vdots \\ \vdots & & \ddots & \vdots \\ -ha_{s1}D\mathbf{f}(\mathbf{y}_0) & \cdots & -ha_{s,s-1}D\mathbf{f}(\mathbf{y}_0) & \mathbf{I} - ha_{ss}D\mathbf{f}(\mathbf{y}_0) \end{pmatrix} .$$

▶ Simplified Newton iteration

$$\mathfrak{g}^{(0)} = 0 \quad , \quad \mathfrak{g}^{(k+1)} = \mathfrak{g}^{(k)} - DF(0)^{-1}F(\mathfrak{g}^{(k)}) \; , \quad k = 0, 1, 2, \dots \; .$$

Retrieval of increments $\mathbf{k}_i$ from $\mathbf{g}_i$: consider l^{th} component, $l = 1, \dots, d$. With $\mathbf{g}_i = (g_{i,1}, \dots, g_{i,d})^T \in \mathbb{R}^d$

$$g_{i,l} = h \sum_{j=1}^{s} a_{ij} k_{j,l} \quad \Longleftrightarrow \quad \left(g_{i,l}\right)_{l=1}^{d} = h\mathfrak{A} \left(k_{i,l}\right)_{l=1}^{d} \; .$$

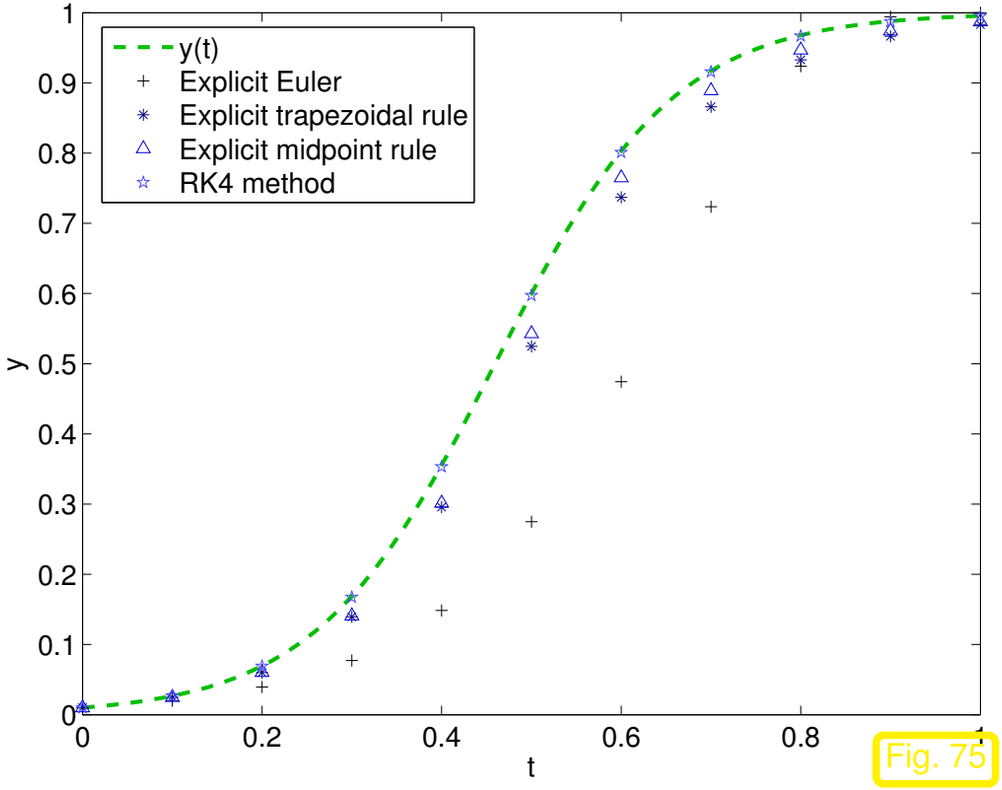
$\mathfrak{A}$ regular ➢ $\mathbf{k}_i$ by solving s linear systems of equations with coefficient matrix $\mathfrak{A}$.

The simplified Newton method can of course also be applied to the standard form of the increment equations from Def. 2.3.5.

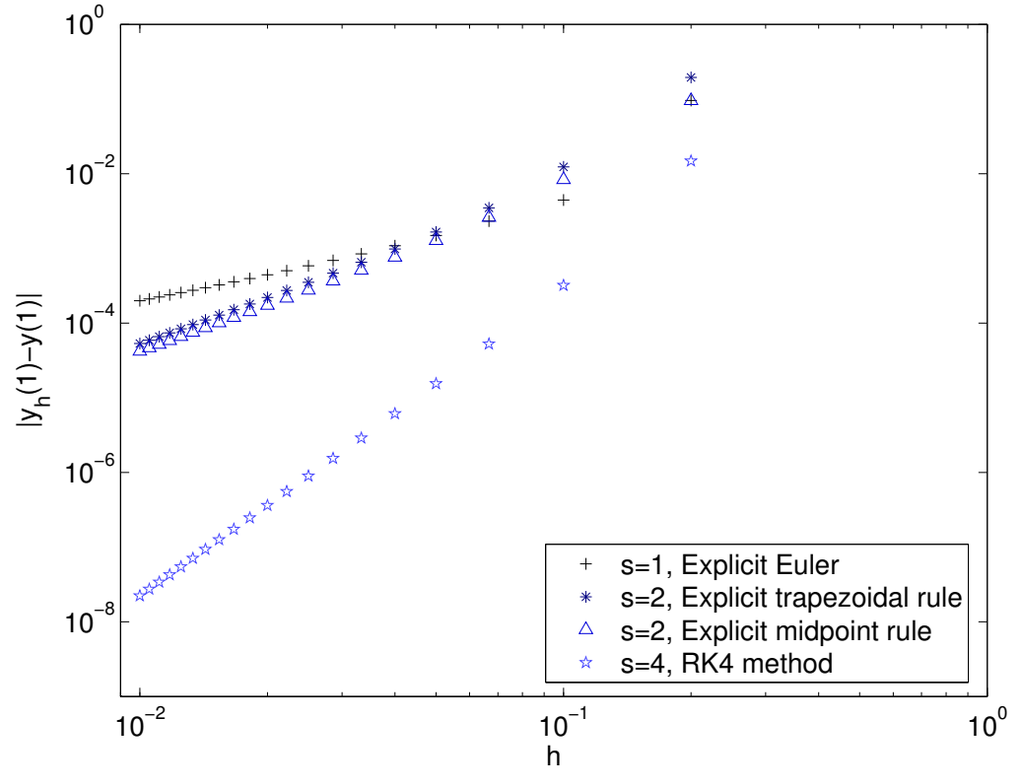
2.3.2 Convergence

Example 2.3.22 (Convergence of explicit Runge-Kutta methods).

- Scalar logistic differential equations (1.2.2), $\lambda = 10$, $y(0) = 0.01$, $T = 1$
- Explicit (single step) Runge-Kutta method, uniform time steps h



$y_h(j/10)$, $j = 1, \ldots, 10$ for explicit RK methods



Convergence of the error $y_h(1) - y(1)$ ◇

P. Grohs

rev 63606, February 17, 2014

Many of our results concerning collocation methods (→ Sect. 2.2) continue to be true for the more general class of Runge-Kutta methods (with essentially the same proofs):

Lemmas 2.2.7, 2.2.13 are valid for (single step) Runge-Kutta methods from Def. 2.3.5

Lemma 2.3.23 (Consistency von (single step) Runge-Kutta methods)**.**
Under the assumptions of Lemma 2.2.7, a (single step) Runge-Kutta method (→ Def. 2.3.5) is consistent (→ Def. 2.1.8) if and only if $\sum_{i=1}^{s} b_i = 1$.

Question: When is a Runge-Kutta method consistent (⇒ convergent, Thm. 2.1.19) of order p **?**

$$\Updownarrow$$

(consistency) conditions for coefficients a_{ij}, b_i

Auxiliary tool to determine conditions : *Taylor expansion* (of the consistency error $\boldsymbol{\tau}(t, \mathbf{y}, h)$ → Def. 2.1.11)

Assumption: $\mathbf{f}$ "sufficiently smooth"

Example 2.3.24 (RK conditions for order of consistency). $p = 3$] [12, Sect. 4.2.2]

Consistency error: $\boldsymbol{\tau}(t,\mathbf{y},h) := (\boldsymbol{\Phi}^{t,t+h} - \boldsymbol{\Psi}^{t,t+h})\mathbf{y}$ (h sufficiently small); .

Focus: autonomous differential equation $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\mathbf{f}$ "sufficiently smooth"

Fix initial value $\mathbf{y}_0 \in D$, WLOG $t_0 = 0$ (cmp. Rem. 1.1.15)

❶ Taylor expansion of the continuous evolution in h around $h = 0$:

$$\boldsymbol{\Phi}^h \mathbf{y}_0 = \mathbf{y}(h) = \mathbf{y}_0 + \dot{\mathbf{y}}(0)h + \tfrac{1}{2}\ddot{\mathbf{y}}(0)h^2 + \tfrac{1}{6}\mathbf{y}^{(3)}(0)h^3 + O(h^4) \;, \tag{2.3.25}$$

with

$$\dot{\mathbf{y}}(0) = \mathbf{f}(\mathbf{y}_0) \;,$$

$$\ddot{\mathbf{y}}(0) = D\mathbf{f}(\mathbf{y}_0)\dot{\mathbf{y}}(0) = D\mathbf{f}(\mathbf{y}_0)\mathbf{f}(\mathbf{y}_0) \;,$$

$$\mathbf{y}^{(3)}(0) = D^2\mathbf{f}(\mathbf{y}_0)(\dot{\mathbf{y}}(0),\dot{\mathbf{y}}(0)) + D\mathbf{f}(\mathbf{y}_0)\ddot{\mathbf{y}}(0) = D^2\mathbf{f}(\mathbf{y}_0)(\mathbf{f}(\mathbf{y}_0),\mathbf{f}(\mathbf{y}_0)) + D\mathbf{f}(\mathbf{y}_0)D\mathbf{f}(\mathbf{y}_0)\mathbf{f}(\mathbf{y}_0) \;.$$

for analogous considerations compare Ex. 2.1.15.

$$\boxed{\begin{aligned}\mathbf{\Phi}^h\mathbf{y}_0 = \; & \mathbf{y}_0 + h\mathbf{f}(\mathbf{y}_0) + \tfrac{1}{2}h^2 D\mathbf{f}(\mathbf{y}_0)\mathbf{f}(\mathbf{y}_0) + \\ & \tfrac{1}{6}h^3(D\mathbf{f}(\mathbf{y}_0)D\mathbf{f}(\mathbf{y}_0)\mathbf{f}(\mathbf{y}_0) + D^2\mathbf{f}(\mathbf{y}_0)(\mathbf{f}(\mathbf{y}_0),\mathbf{f}(\mathbf{y}_0))) + O(h^4)\;.\end{aligned}} \tag{2.3.26}$$

❷ Taylor expansion of the discrete evolution in h around $h = 0$

$$\Updownarrow$$

Taylor expansion of the increments $\mathbf{k}_i$ in h around $h = 0$

$$\begin{aligned}\mathbf{k}_i &= \mathbf{f}(\mathbf{y}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j) = \\ &= \mathbf{f}(\mathbf{y}_0) + h\,D\mathbf{f}(\mathbf{y}_0)\left(\sum_{j=1}^{s} a_{ij}\mathbf{k}_j\right) + \tfrac{1}{2}h^2 D^2\mathbf{f}(\mathbf{y}_0)\left(\sum_{j=1}^{s} a_{ij}\mathbf{k}_j, \sum_{j=1}^{s} a_{ij}\mathbf{k}_j\right) + O(h^3)\end{aligned} \tag{2.3.27}$$

Substitute “shorter Taylor expansions” for the increments

increments are multiplied by h !

$$
\begin{aligned}
\mathbf{k}_i =& \mathbf{f}(\mathbf{y}_0) + hD\mathbf{f}(\mathbf{y}_0)\sum_{j=1}^{s} a_{ij}\left(f\big(\mathbf{y}_0 + hD\mathbf{f}(\mathbf{y}_0)\sum_{l=1}^{s} a_{il}\mathbf{k}_l + O(h^2)\big)\right) + \\
& \tfrac{1}{2}h^2 D^2\mathbf{f}(\mathbf{y}_0)\left(\sum_{j=1}^{s} a_{ij}(\mathbf{f}(\mathbf{y}_0) + O(h)), \sum_{j=1}^{s} a_{ij}(\mathbf{f}(\mathbf{y}_0) + O(h))\right) + O(h^3) \\
=& \mathbf{f}(\mathbf{y}_0) + h \underbrace{\cdot \sum_{j=1}^{s} a_{ij}}_{=c_i \text{ , compare Rem. 2.3.15}} D\mathbf{f}(\mathbf{y}_0)\mathbf{f}(\mathbf{y}_0) + \\
& h^2\big(\sum_{l=1}^{s} a_{il}c_l\big)D\mathbf{f}(\mathbf{y}_0)D\mathbf{f}(\mathbf{y}_0)\mathbf{f}(\mathbf{y}_0) + h^2\tfrac{1}{2}c_i^2 D^2\mathbf{f}(\mathbf{y}_0)(\mathbf{f}(\mathbf{y}_0), \mathbf{f}(\mathbf{y}_0)) + O(h^3) \;.
\end{aligned}
$$

Note: $\mathbf{\Psi}^h\mathbf{y}_0 = \mathbf{y}_0 + h\sum_{i=1}^{s} b_i\mathbf{k}_i$ ➤ expansion up to $O(h^3)$ sufficient

$$\mathbf{\Psi}^h\mathbf{y}_0 = \mathbf{y}_0 + \left(h\sum_{i=1}^{s} b_i\right)\mathbf{f}(\mathbf{y}_0) + \left(h^2\sum_{i=1}^{s} b_i c_i\right) D\mathbf{f}(\mathbf{y}_0)\mathbf{f}(\mathbf{y}_0) + \left(h^3\sum_{i=1}^{s} b_i \sum_{j=1}^{s} a_{ij}c_j\right) D\mathbf{f}(\mathbf{y}_0)D\mathbf{f}(\mathbf{y}_0)\mathbf{f}(\mathbf{y}_0) + \left(\tfrac{1}{2}h^3\sum_{i=1}^{s} b_i c_i^2\right) D^2\mathbf{f}(\mathbf{f}(\mathbf{y}_0),\mathbf{f}(\mathbf{y}_0)) + O(h^4)\,. \tag{2.3.28}$$

❸ Set the coefficients of the *linearly independent* elementary differential

$$1, \mathbf{f}(\mathbf{y}_0), D\mathbf{f}(\mathbf{y}_0)\mathbf{f}(\mathbf{y}_0), D\mathbf{f}(\mathbf{y}_0)D\mathbf{f}(\mathbf{y}_0)\mathbf{f}(\mathbf{y}_0), D^2\mathbf{f}(\mathbf{y}_0)(\mathbf{f}(\mathbf{y}_0),\mathbf{f}(\mathbf{y}_0))$$

in (2.3.28) and (2.3.26) to be equal

Sufficient & *necessary* conditions for order of consistency $p = 3$ of a RK method (that is invariant w.r.t. autonomization $\rightarrow$ Rem. 2.3.15) :

$$\sum_{i=1}^{s} b_i = 1\,, \tag{2.3.29}$$

$$\sum_{i=1}^{s} b_i c_i = \frac{1}{2}\,, \tag{2.3.30}$$

$$\begin{aligned} \sum_{i=1}^{s} b_i c_i^2 &= \frac{1}{3}\,, \\ \sum_{i=1}^{s} b_i \sum_{j=1}^{s} a_{ij} c_j &= \frac{1}{6}\,. \end{aligned} \tag{2.3.31}$$

☞ (2.3.29) sufficient **&** necessary for order of consistency $p = 1$, compare Lemma 2.3.23

☞ (2.3.29) **+** (2.3.30) sufficient **&** necessary for order of consistency $p = 2$

◇

Remark 2.3.32 (Butcher trees).

General combinatorial algorithm to determine the RK consistency conditions: Butcher trees [12, Sect. 4.2.3], [21, Ch. III]

△

→ Construction of RK methods of given order of convergence by solving the (nonlinear) conditions (of the type (2.3.29)-(2.3.31)):

p	1	2	3	4	5	6	7	8	9	10	20
♯B.G.	1	2	4	8	17	37	85	200	486	1205	20247374

Some orders of convergence of Runge-Kutta methods:

Explicit methods		Implicit methods	
Explicit Euler method (2.2.1)	$p=1$		
Explicit trapezoidal rule (2.3.3)	$p=2$	Implicit Euler method (2.2.1)	$p=1$
Explicit midpoint rule (2.3.4)	$p=2$	Implicit midpoint rule (2.2.19)	$p=2$
Classic Runge-Kutta method (2.3.11)	$p=4$	Gauss collocation method	$p=2s$
Kuttas $3/8$-rule (2.3.12)	$p=4$		

Many more RK methods ▷ [22, 23]

Order barriers:

for explicit Runge-Kutta methods $p \leq s$

for general Runge-Kutta methods $p \leq 2s$

➢ Gauss collocation methods achieve maximum order

Remark 2.3.33 ("Butcher barriers" for explicit RK SSM).

Order p	1	2	3	4	5	6	7	8	≥ 9
Minimum number of stages s	1	2	3	4	6	7	9	11	$\geq p+3$

A more general formula for the minimum number of stages has not been deduced yet.

△

P. Grohs

rev 63606, February 17, 2014

Remark 2.3.34 (Why single step methods of high order?).

The general convergence theory for single step methods from section 2.1.3 provides for sufficient smoothness of the solution of the IVP the asymptotic error estimate

$$\operatorname{err}(\mathcal{G}) := \max_{k=1,\dots,N} \|\mathbf{y}_k - \mathbf{y}(t_k)\| \leq C h_{\mathcal{G}}^p \quad \text{for } h_{\mathcal{G}} \text{ sufficiently small,} \tag{2.3.35}$$

compare Thm. 2.1.19, where the constant $C > 0$ does not depend on the maximum time step size $h_{\mathcal{G}} > 0$, but is also typically unknown.

Therefore, (2.3.35) does in general not provide information about the integration error on a specific time grid ($\rightarrow$ discussion at the end of section 2.1.1), or about the time step size necessary to guarantee a certain accuracy.

As mentioned earlier, under the assumption that the estimate (2.3.35) is sharp, we can only deduce

which reduction of the integration error is expected for refining the step size.

Assumption: Estimate (2.3.35) is sharp

Then, we can predict how much more accurate the solution will be for increasing the computational work in the numerical integration of an IVP.

Convention: $\boxed{\text{Computational work} \quad \sim \quad \text{total number of evaluations of } \mathbf{f}}$

➢ for s-stage Runge-Kutta method ($\rightarrow$ Def. 2.3.5):

Computational work $\sim$ $s\cdot$number of (steps) $\sim$ Csh^{-1} with a constant $C > 0$

for uniform time grid, step size $h > 0$

Goal: specified reduction of the error: for $0 < \rho < 1$

$$\frac{\mathrm{err}(\mathcal{G}_{\mathrm{neu}})}{\mathrm{err}(\mathcal{G}_{\mathrm{alt}})} \stackrel{!}{=} \rho \quad \stackrel{(2.3.35)}{\Longrightarrow} \quad \frac{h_{\mathrm{neu}}^p}{h_{\mathrm{alt}}^p} \stackrel{!}{=} \rho \quad \Leftrightarrow \quad \boxed{h_{\mathrm{neu}} = \rho^{1/p} h_{\mathrm{alt}}} \; .$$

Rule of thumb for a RK SSM of (consistency = convergence) order $p \in \mathbb{N}$ (uniform step size)

Increase of the computational work by factor $\rho^{-1/p}$ ► Reduction of the error by factor ρ

☞ The higher the order, the less relative additional work is necessary to reduce the error by a specified factor.

2.4 Extrapolation methods [12, Sect. 4.3]

2.4.1 The combination trick

Single step method for initial value problem

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}) \quad , \quad \mathbf{y}(t_0) = \mathbf{y}_0 \; , \quad (t_0, \mathbf{y}_0) \in \Omega \; , \qquad \text{((1.1.13))}$$

considered on $[t_0, T]$, produces grid function $\mathbf{y}_{\mathcal{G}} = (\mathbf{y}_k)_{k=1}^N$ as solution: $\mathbf{y}_k \approx \mathbf{y}(t_k)$, $t_N = T$.

For equidistant step size $h > 0$, we also write $\mathbf{y}_h(t_k) := \mathbf{y}_k$, i.e., $\mathbf{y}_h(T) = \mathbf{y}_N$.

Sect. 2.1.3: single step method for $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, convergent of order p

$$\exists C > 0: \quad \|\mathbf{y}_h(T) - \mathbf{y}(T)\| \leq C h^p \quad \text{for } h \to 0,\; N = T/h \in \mathbb{N} \; .$$

(Assumption: Equidistant time steps of length $h > 0$)

Speculation: $\exists \mathbf{c} \in \mathbb{R}^d$: $\boxed{\mathbf{y}_h(T) - \mathbf{y}(T) = \mathbf{c}h^p + O(h^{p+1}) \quad \text{for } h \to 0}$. (2.4.1)

$$\begin{aligned} \mathbf{y}_h(T) - \mathbf{y}(T) &= \mathbf{c}h^p + O(h^{p+1}) && \text{(I)} \\ \mathbf{y}_{h/2}(T) - \mathbf{y}(T) &= \mathbf{c}2^{-p}h^p + O(h^{p+1}) && \text{(II)} \end{aligned}$$

(I)-2^p·(II): $$y_h(T) - 2^p y_{h/2}(T) - (1-2^p)\mathbf{y}(T) = O(h^{p+1}) ,$$

$$\Rightarrow \quad \frac{y_h(T) - 2^p y_{h/2}(T)}{1-2^p} - \mathbf{y}(T) = O(h^{p+1}) .$$

combined method, convergent of order $p+1$ **!**

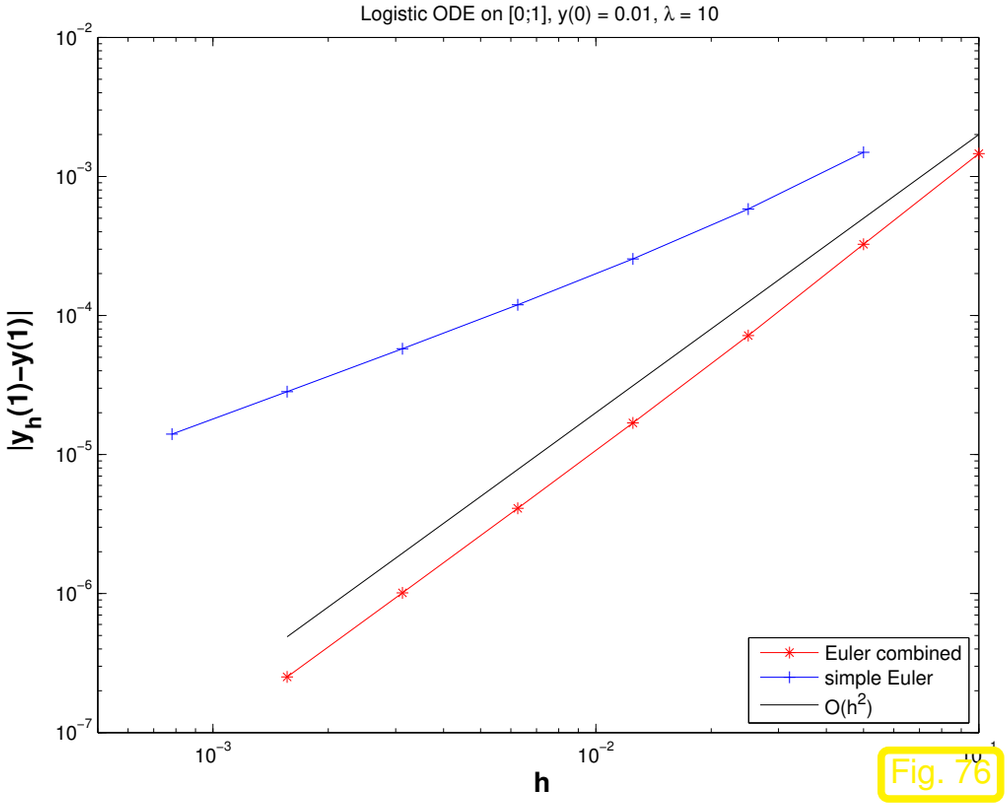
Example 2.4.2 (Convergence of combined methods).

IVP for logistic differential equation (2.2.57): ($\to$ Ex. 1.2.1)

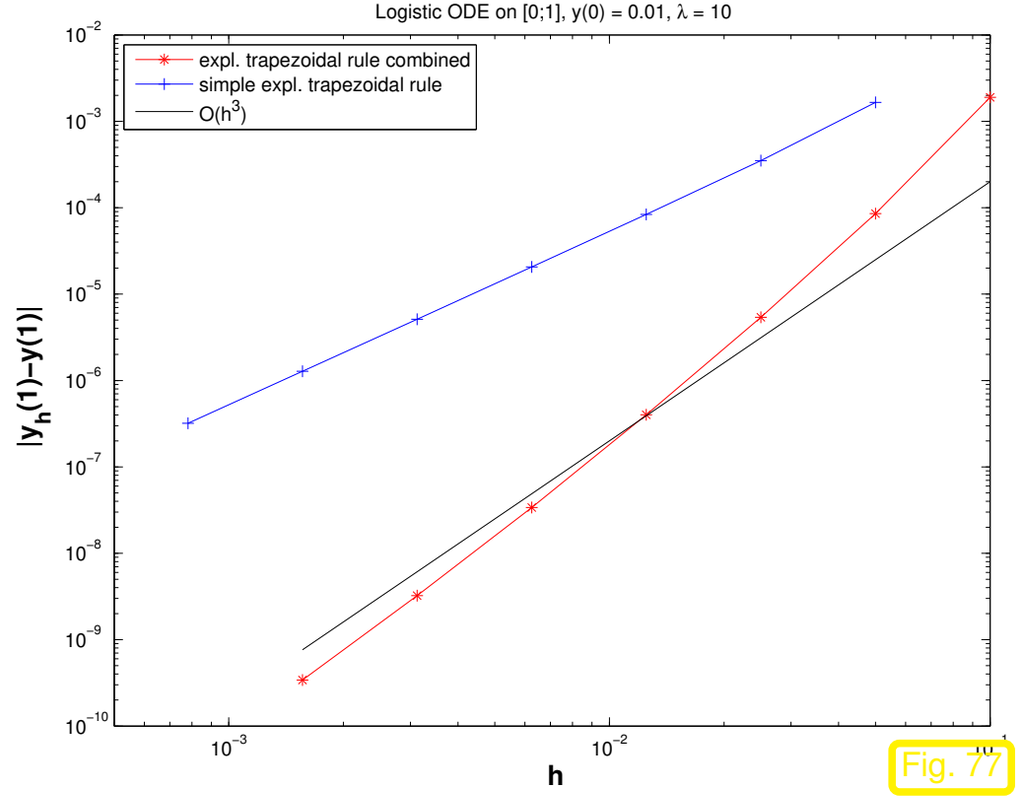
$$\dot{y} = \lambda y(1-y) , \quad y_0 = 0.01 \quad \Rightarrow \quad y(t) = \frac{1}{1 + 99 \cdot e^{-\lambda t}} , \quad t \in \mathbb{R} .$$

Base methods: explicit Euler method (1.4.2), $p = 1$

explicit trapezoidal rule (2.3.3), $p = 2$



Euler method



Explicit trapezoidal rule ◇

P. Grohs

rev 63606, February 17, 2014

Note: If $h \mapsto \mathbf{y}_h(T)$ smooth, method convergent of order p, then (2.4.1) is obvious: Taylor expansion !

This will be discussed more extensively in section 2.4.3.

2.4.2 Extrapolation idea

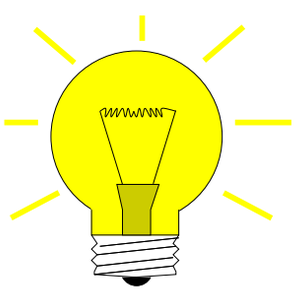
Reminder ($\rightarrow$ course "Numerical methods"): Romberg quadrature ($\rightarrow$ [13, Sect. 9.4])

Abstract frame:

Problem: $\mathbf{\Pi} : X \mapsto \mathbb{R}^d$, we look for $\mathbf{\Pi}(x_0)$ for fixed $x_0 \in X$, $X \hat{=}$ data space

Family of numerical methods $\left\{\mathbf{\Pi}_h : X \mapsto \mathbb{R}^d\right\}_h$ ➤ approximations $\mathbf{\Pi}_h(x_o) \approx \mathbf{\Pi}(x_0)$

$\mathbf{\Pi}_h$ depends on scalar discretization parameter $h > 0$ (e.g., time step)



- Compute $\mathbf{\Pi}_h(x_0)$ for $h \in \{h_1, \dots, h_k\}$ ("series of step sizes", $h_i > h_{i+1}$)
- Compute interpolation polynomial $\mathbf{p} \in (\mathcal{P}_{k-1})^d$ with
$$\mathbf{p}(h_i) = \mathbf{\Pi}_{h_i}(x_0), \quad i = 1, \dots, k \quad .$$
- Better (?) approximation $\boxed{\mathbf{\Pi}(x_0) \approx \mathbf{p}(0)}$

Example 2.4.3 (Romberg quadrature).

Interpretation of the abstract frame for the Romberg quadrature: $X = C^0([a,b])$, $a,b \in \mathbb{R}$, $a < b$

$$\mathbf{\Pi}(f) := \int_a^b f(x)\,\mathrm{d}x \quad , \quad \mathbf{\Pi}_h := \tfrac{h}{2} f(a) + h \sum_{j=1}^{N-1} f(a + j\tfrac{b-a}{N}) + \tfrac{h}{2} f(b) \; , \quad h := \frac{1}{N} \; ,$$

$\mathbf{\Pi}_h \mathrel{\hat{=}}$ trapezoidal rule for numerical quadrature

Discretization parameter $h = \frac{1}{N}$, $N \in \mathbb{N}$ (“mesh width” of the trapezoidal rule): can only assume discrete values !

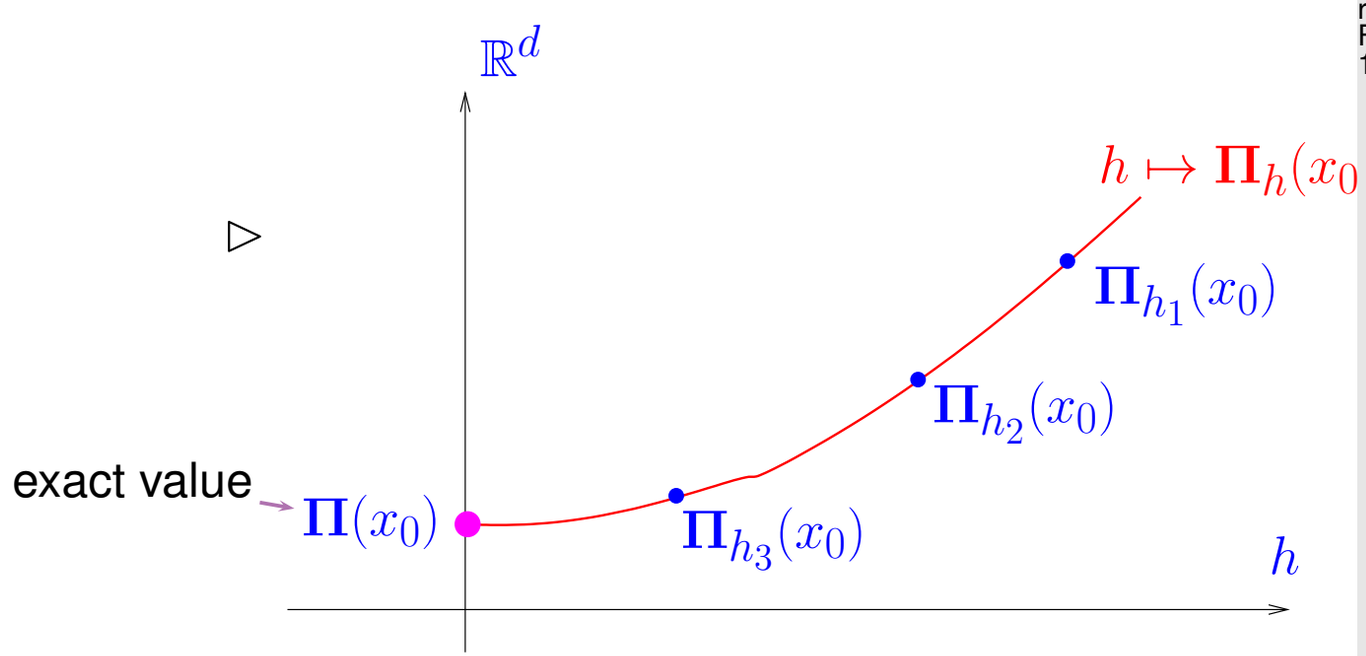
◇

P. Grohs

rev 63606, February 17, 2014

Visualization: ▷

Idea of extrapolation method



Remark 2.4.4 (Scale invariance of extrapolation).

$p(t) \in \mathcal{P}_{k-1} \hat{=}$ Interpolation polynomial for $(t_1, y_1), \dots, (t_k, y_k)$
$\widetilde{p}(t) \in \mathcal{P}_{k-1} \hat{=}$ Interpolation polynomial for $(\xi t_1, y_1), \dots, (\xi t_k, y_k)$ for a $\xi \in \mathbb{R}$

$$p(0) = \widetilde{p}(0)$$

(If $p(t) = \sum_{j=0}^{s} a_j t^j$, then all polynomials $p_\xi(t) = \sum_{j=0}^{s} a_j (\xi t)^j$ have obviously the same value for $t = 0$.)

It suffices to specify the quotients $\eta_i := \dfrac{h_1}{h_i}$!

△

P. Grohs

rev 63606, February 17, 2014

Algorithm: Aitken-Neville scheme [13, Sect. 9.4] $\rightarrow$ course “Numerical methods”

Recursive computation of values of interpolation polynomials for $h = 0$, $p = 1$:

$$T_{i1} := \mathbf{\Pi}_{h_i}(x_0)\,, \quad i = 1, \dots, k\,, \tag{2.4.5}$$

$$T_{il} := T_{i,l-1} + \frac{T_{i,l-1} - T_{i-1,l-1}}{\dfrac{h_{i-l+1}}{h_i} - 1}\,, \quad 2 \le l \le k\,. \tag{2.4.6}$$

$$\begin{array}{ccccccc}
T_{11} & & & & & & \\
 & \searrow & & & & & \\
T_{21} & \to & T_{22} & & & & \\
\vdots & & & \ddots & & & \\
T_{k-1,1} & \to & \cdots & \to & T_{k-1,k-1} & & \\
 & \searrow & & \searrow & & \searrow & \\
T_{k1} & \to & \cdots & \to & T_{k,k-1} & \to & T_{kk}
\end{array}$$

Extrapolation tableau ▷

MATLAB-CODE : Aitken-Neville extrapolation

```
function T = anexpol(y,h)
k = length(h);
T(1) = y(1);
for i=2:k
  T(i) = y(i);
  for l=i-1:-1:1
    T(l)=T(l+1)+(T(l+1)-T(l))/...
          (h(l)/h(i)-1);
  end
end
```

$\eta_l : \eta_i$

$$\begin{array}{cccccc}
\mathtt{T(1)} & & \mathtt{T(2)} & & \mathtt{T(3)} & \cdots \ \mathtt{T(k)} \\
\hline
T_{11} = y_1 & & & & & \\
\downarrow & & & & & \\
T_{22} & \leftarrow & T_{21} = y_2 & & & \\
\downarrow & & \downarrow & & & \\
T_{33} & \leftarrow & T_{32} & \leftarrow & T_{31} = y_3 & \\
\downarrow & & \downarrow & & \downarrow & \\
\vdots & & \vdots & & \vdots & \\
\downarrow & & \downarrow & & \downarrow & \\
T_{kk} & \leftarrow & T_{k,k-1} & \leftarrow & \cdots & \leftarrow \ T_{k,1}
\end{array}$$

Output: lowest row of tableau decreasing

P. Grohs

rev 63606, February 17, 2014

☞ Extrapolation "works", if

- $\lim_{h\to 0} \mathbf{\Pi}_h(x_0) = \mathbf{\Pi}(x_0) \quad \hat{=} \quad$ convergence,
- $h \mapsto \mathbf{\Pi}_h(x_0)$ "behaves like a polynomial for small h."

Definition 2.4.7 ((Truncated) asymptotic expansion)**.**
$h \mapsto \mathbf{\Pi}_h(x_0)$ ($x_0 \in X$ fixed) has a (truncated) asymptotic expansion in h up to the order k, if there exist constants$^{(*)}$ $\alpha_0, \alpha_1, \dots, \alpha_k \in \mathbb{R}^d$ and a function $h \mapsto R_k(h)$ that is uniformly bounded for sufficiently small h such that

$$\mathbf{\Pi}_h(x_0) = \alpha_0 + \alpha_1 h + \alpha_2 h^2 + \cdots + \alpha_k h^k + R_k(h) h^{k+1} \quad \text{for small } h > 0 \ .$$

$(*)$ α_i constants $\hat{=}$ α_i independent of h !

Obvious: Sufficient **&** necessary condition for convergence: $\alpha_0 = \mathbf{\Pi}(x_0)$

Theorem 2.4.8 (Convergence of extrapolated values)**.**
Let $\mathbf{\Pi}_h(x_0)$ have an asymptotic expansion in h up to the order k according to Def. 2.4.7. Then, the values from the extrapolation tableau, compare (2.4.5), (2.4.6)*, satisfy for sufficiently small $h_j > 0$*

$$\left\| T_{i,l} - \alpha_0 \right\| \leq \|\alpha_l\| \, h_{i-l+1} \cdot \cdots \cdot h_i + C \cdot \sum_{j=i-l+1}^{i} \|R_k(h_j)\| \, h_j^{l+1} \, , \quad 1 \leq i, l \leq k \, ,$$

where $C > 0$ <u>only</u> depends on the ratios $h_i : h_j$.

P. Grohs

rev 63606, February 17, 2014

Proof: Every $T_{i,k}$ in the extrapolation tableau can be interpreted as "final value" T_{kk} of a portion of the tableau ➥ It suffices to show the claim for $i = l = k$

Prerequisite: Existence of a (truncated) asymptotic expansion $\rightarrow$ Def. 2.4.7

$$T_{i,1} = \mathbf{\Pi}_{h_i}(x_0) = \alpha_0 + \alpha_1 h_i + \alpha_2 h_i^2 + \cdots + \alpha_k h_i^k + R_k(h) h_i^{k+1} \quad \text{for small } h_i > 0 \, .$$

Extrapolation polynomial for $(h_i, T_{i,1})$, $i = 1, \ldots, k$: $q \in \mathcal{P}_{k-1}$, represented by Lagrange polynomial, compare (2.2.2)

$$q(t) = \sum_{i=1}^{k} T_{i,1} L_i(t) \, , \quad L_i \in \mathcal{P}_{k-1} \, , \quad L_i(h_j) = \delta_{ij}, \; i, j = 1, \ldots, k \, .$$

$$\sum_{i=1}^{k} L_i(0)h_i^j = \begin{cases} 1 & \text{for } j = 0 \ , \\ 0 & \text{for } 1 \le j \le k-1 \ , \\ (-1)^{k-1} h_1 \cdot \cdots \cdot h_k & \text{for } j = k \ . \end{cases} \tag{2.4.9}$$

Proof of (2.4.9): for $0 \le j \le k-1$, $r_j(t) := \sum_{i=1}^{k} h_i^j L_i(t) \in \mathcal{P}_{k-1}$ coincides with $t \mapsto t^j$. For $j = k$, $t^k - r_k(t) \in \mathcal{P}_k$ has the roots h_i, $i = 1, \dots, k$ and leading coefficients 1:

$$\blacktriangleright \quad t^k - r_k(t) = (t - h_1) \cdot \cdots \cdot (t - h_k) \ .$$

This implies (2.4.9).

$$\begin{aligned} T_{k,k} = q(0) &= \sum_{i=1}^{k} L_i(0) T_{i,1} = \sum_{i=1}^{k} L_i(0) \left(\sum_{j=0}^{k} \alpha_j h_i^j + R_k(h_i) h_i^{k+1} \right) \\ &= \sum_{j=0}^{k} \alpha_j \sum_{i=1}^{k} h_i^j L_i(0) + \sum_{i=1}^{k} L_i(0) R_k(h_i) h_i^{k+1} \\ &\overset{(2.4.9)}{=} \alpha_0 + \alpha_k \cdot (-1)^{k-1} h_1 \cdot \cdots \cdot h_k + \sum_{i=1}^{k} L_i(0) R_k(h_i) h_i^{k+1} \ . \end{aligned}$$

Therefore, the claim holds with $C := \max_{i=1,\dots,l} |L_i(0)|$. Note that $L_i(0)$ only depends on the ratios $h_i : h_j$, compare Rem. 2.4.4. $\square$

2.4.3 Extrapolation of single step methods

Initial value problem (1.1.13): $\dot{\mathbf{y}} = f(t,\mathbf{y})$, $\mathbf{y}(t_0) = \mathbf{y}_0$ ➢ solution $t \mapsto \mathbf{y}(t)$

Consider the initial value problem on the fixed time interval $[t_0, T]$

Assumption: f "'"sufficiently" smooth $\Leftrightarrow$ $\mathbf{y}(t)$ "'"sufficiently" smooth

Given: consistent single step method $\leftrightarrow$ discrete evolution ($\rightarrow$ Lemma 2.1.9)

$$\boldsymbol{\Psi}^{t,t+h}\mathbf{y} = \mathbf{y} + h\boldsymbol{\psi}(t,\mathbf{y},h) \quad , \quad \boldsymbol{\psi}(t,\mathbf{y},0) = f(t,\mathbf{y}) \; , \quad (t,\mathbf{y}) \in \Omega, \; h \text{ small} \; . \tag{2.4.10}$$

Assumption:
- increment function $\boldsymbol{\psi}$ continuously differentiable in $(t,\mathbf{y})$
- SSM has $\boxed{\text{order of consistency = order of convergence } p \in \mathbb{N}}$ ($\rightarrow$ Thm. 2.1.19)

Given: final time $T \in J(t_0,\mathbf{y}_0)$ ➡ uniform time steps $h = (T-t_0)/N$, $N \in \mathbb{N}$

Single step method ➡ grid function $\{\mathbf{y}_k\}_{k=0}^N$, $\mathbf{y}_N \approx \mathbf{y}(T)$

Theorem 2.4.11 (Asymptotic expansion of the discretization error of a SSM)**.**
There exists a $K \in \mathbb{N}$ *(dependent on the smoothness of* $\mathbf{f}$*) and smooth functions* $\mathbf{e}_i : J(t_0, \mathbf{y}_0) \mapsto \mathbb{R}^d$, $i = p, p+1, \dots, p+K$, *with* $\mathbf{e}_i(0) = 0$ *and (for sufficiently small* h*) uniformly bounded functions* $(T,h) \mapsto \mathbf{r}_{k+p+1}(T,h)$, $0 \le k \le K$ *such that*

$$\mathbf{y}_N - \mathbf{y}(T) = \sum_{l=0}^{k} \mathbf{e}_{l+p}(T) h^{l+p} + \mathbf{r}_{k+p+1}(T,h) h^{k+p+1} \quad \textit{for small } h \ .$$

Hereby, there holds $\left\| \mathbf{r}_{k+p+1}(T,h) \right\| = O(T - t_0)$ for $T - t_0 \to 0$ uniform in $h < T$,
$\|\mathbf{e}(T)\| = O(T - t_0)$ for $T - t_0 \to 0$.

P. Grohs

rev 63606, February 17, 2014

Proof. Assumption: $\mathbf{f}$, $\mathbf{y}(t)$ "'sufficiently" smooth

Furthermore, we assume the *global Lipschitz continuity* of the increment function $\boldsymbol{\psi}$ of the SSM from (2.4.10):

$$\exists L > 0: \quad \|\boldsymbol{\psi}(t, \mathbf{z}, h) - \boldsymbol{\psi}(t, \mathbf{w}, h)\| \le L \, \|\mathbf{z} - \mathbf{w}\| \quad \text{uniform in } t_0 \le t \le T, \, h \ . \tag{2.4.12}$$

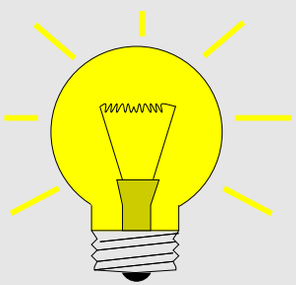
(Compactness arguments, cmp. proof of Thm. 2.1.19, make it possible to waive this assumption.)

Consequence of the order of consistency p ($\rightarrow$ Def. 2.1.13) and of the smoothness of $\mathbf{f}$: for consistency errors ($\rightarrow$ Def. 2.1.11) along the solution trajectory holds (only there the estimate for the consistency error is used in the proof of Thm. 2.1.19 !)

$$\boldsymbol{\tau}(t,\mathbf{y}(t),h) := \mathbf{y}(t+h) - \boldsymbol{\Psi}^{t,t+h}\mathbf{y}(t) = \mathbf{d}(t)h^{p+1} + O(h^{p+2}) \quad \text{for } h \to 0 \;, \tag{2.4.13}$$

with continuous function $\mathbf{d} : [t_0,T] \mapsto \mathbb{R}^d$. This is a result of Taylor expansion, compare Ex. 2.3.24:

RK SSM: $\mathbf{d}$ only depends on the derivatives of $\mathbf{f}$ ➢ $\mathbf{d}$ "sufficiently smooth"



Idea: Consider SSM with modified increment function

$$\widehat{\boldsymbol{\psi}}(t,\mathbf{u},h) := \boldsymbol{\psi}(t,\mathbf{u}+\mathbf{e}(t)h^p,h) - (\mathbf{e}(t+h)-\mathbf{e}(t))h^{p-1} \;, \tag{2.4.14}$$

with "sufficiently smooth" function $\mathbf{e} : [t_0,T] \mapsto \mathbb{R}^d$.

Note: $\widehat{\boldsymbol{\psi}}$ satisfies (2.4.12) with the same $L > 0$, too.

Why do we consider this modified SSM **?**

$\mathbf{y}_j$/$\widehat{\mathbf{y}}_j$, $j = 0,\dots,N \mathrel{\hat{=}}$ grid functions generated by original/modified SSM with step size $h := \frac{(T-t_0)}{N}$.
Set $\widehat{\mathbf{y}}_0 = \mathbf{y}_0$

$$\blacktriangleright \quad \widehat{\mathbf{y}}_j = \mathbf{y}_j - \mathbf{e}(t_j)h^p \;, \quad t_j := t_0 + jh \;, \quad j = 0,\dots,N \;. \tag{2.4.15}$$

Proof von (2.4.15) by induction:

$$\begin{aligned}\widehat{\mathbf{y}}_{j+1} &= \widehat{\mathbf{y}}_j + h\widehat{\boldsymbol{\psi}}(t_j, \widehat{\mathbf{y}}_j, h) \\ &\overset{(2.4.14)}{=} \widehat{\mathbf{y}}_j + h\boldsymbol{\psi}(t_j, \widehat{\mathbf{y}}_j + \mathbf{e}(t_j)h^p, h) - h^p(\mathbf{e}(t_{j+1}) - \mathbf{e}(t_j)) \\ &\overset{(*)}{=} \underbrace{\mathbf{y}_j + h\boldsymbol{\psi}(t_j, \mathbf{y}_j, h)}_{=\mathbf{y}_{j+1}} - \mathbf{e}(t_{j+1})h^p \ .\end{aligned}$$

$(*)$ $\leftarrow$ induction assumption.

P. Grohs

rev 63606, February 17, 2014

Assumption: The modified single step method is consistent with $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$ of order $p+1$

$$\overset{\text{Thm. 2.1.19}}{\Rightarrow} \quad \widehat{\mathbf{y}}_N - \mathbf{y}(T) = \mathbf{r}_{p+1}(T,h)h^{p+1} \ , \quad \left\|\mathbf{r}_{p+1}(T,h)\right\| \leq C \underbrace{\frac{\exp(L(T-t_0)) - 1}{L}}_{=O(T-t_0) \text{ for } T-t_0 \to 0} \ ,$$

with $C > 0$ independent of h, T. $L \hat{=}$ *joint* Lipschitz constant of $\boldsymbol{\psi}$, $\widehat{\boldsymbol{\psi}}$ (w.r.t. $\mathbf{y}$) from (2.4.12)

$$\overset{(2.4.15)}{\Rightarrow} \quad \mathbf{y}_N - \mathbf{y}(T) = \mathbf{e}(T)h^p + \mathbf{r}_{p+1}(T,h)h^{p+1} \ .$$

This gives us the first term in the asymptotic expansion of the theorem.

Inductive application of the argument ➢ Modified SSMs $\widehat{\mathbf{\Psi}}_1 := \widehat{\mathbf{\Psi}}, \widehat{\mathbf{\Psi}}_2, \dots, \widehat{\mathbf{\Psi}}_{k+1}$ consistent to $\dot{\mathbf{y}} = f(t, \mathbf{y})$ of orders $p+1, p+2, \dots, p+k+1$ generate approximations $\widehat{\mathbf{y}}_j^1 := \widehat{\mathbf{y}}_j, \widehat{\mathbf{y}}_j^2, \dots, \widehat{\mathbf{y}}_j^{k+1}$, $j = 1, \dots, N$.

With $\widehat{\mathbf{y}}_k^0 = \mathbf{y}_k$ (telescope sum)

$$\widehat{\mathbf{y}}_j^{l+1} = \widehat{\mathbf{y}}_j^l - \mathbf{e}_l(t_j) h^{p+l} \ , \quad l = 0, \dots, k \ .$$

$$\blacktriangleright \qquad \begin{aligned} \mathbf{y}_N - \mathbf{y}(T) &= \sum_{l=0}^{k} \widehat{\mathbf{y}}_N^l - \widehat{\mathbf{y}}_N^{l+1} + \mathbf{r}_{p+k+1}(T,h) h^{p+k+1} \\ &= \sum_{l=0}^{k} \mathbf{e}_l(T) h^{p+l} + \mathbf{r}_{p+k+1}(T,h) h^{p+k+1} \ . \end{aligned}$$

This implies the claim of the theorem.

? Existence of $\mathbf{e}(t)$ such that the modified SSM has order of consistency $p+1$.

☞ Consider the consistency error of the modified methods **&** Taylor expansion(s)

$$\mathbf{y}(t+h) - \widehat{\mathbf{\Psi}}^{t,t+h} \mathbf{y}(t) = \mathbf{y}(t+h) - \mathbf{y}(t) - h \widehat{\boldsymbol{\psi}}(t, \mathbf{y}(t), h)$$

$$= \mathbf{y}(t+h) - \mathbf{y}(t) - h\boldsymbol{\psi}(t, \mathbf{y}(t) + \mathbf{e}(t)h^p, h) + (\mathbf{e}(t+h) - \mathbf{e}(t))h^p$$

$$= \mathbf{y}(t+h) - \mathbf{y}(t) - h\left(\boldsymbol{\psi}(t, \mathbf{y}(t), h) + \frac{\partial \boldsymbol{\psi}}{\partial \mathbf{y}}(t, \mathbf{y}(t), h)\mathbf{e}(t)h^p + O(h^{2p})\right) + \dot{\mathbf{e}}(t)h^{p+1} + O(h^{p+2})$$

$$\overset{(2.4.13)}{=} \mathbf{d}(t)h^{p+1} + O(h^{p+2}) - \left(\frac{\partial \boldsymbol{\psi}}{\partial \mathbf{y}}(t, \mathbf{y}(t), 0) + \frac{\partial^2 \boldsymbol{\psi}}{\partial \mathbf{y} \partial h}(t, \mathbf{y}(t), 0)h\right)\mathbf{e}(t)h^{p+1}$$

$$+ \dot{\mathbf{e}}(t)h^{p+1} + O(h^{p+2})$$

$$= (\mathbf{d}(t) - \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t, \mathbf{y}(t))\mathbf{e}(t) + \dot{\mathbf{e}}(t))h^{p+1} + O(h^{p+2}) \ .$$

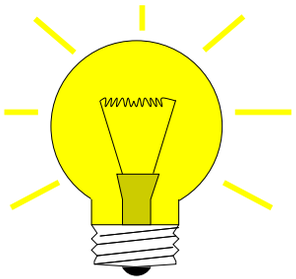
If $\mathbf{e}$ solves the following IVP for an inhomogeneous linear variational equation

$$\dot{\mathbf{e}}(t) = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t, \mathbf{y}(t))\mathbf{e}(t) - \mathbf{d}(t) \quad , \quad \mathbf{e}(0) = 0 \quad \Rightarrow \quad \mathbf{e} \text{ smooth} \ , \tag{2.4.16}$$

then the assumption concerning the modified SSM is satisfied. □

Clear: K depends on the smoothness of $\mathbf{f}$.

Idea: Increase in order by extrapolation ($\rightarrow$ Sect. 2.4.2)



- Choose $N_1 < N_2 < \cdots < N_{k+1}$, $N_i \in \mathbb{N}$
- SSM (step size $h_i = (T-t_0)/N_i$) produces $\mathbf{y}_{N_i}$, $i = 1, \ldots, k+1$
- Polynomial extrapolation $(*)$ from $(h_i, \mathbf{y}_{h_i,N_i})$
  - ➥ approximations $\widetilde{\mathbf{y}}$ with $\|\widetilde{\mathbf{y}} - \mathbf{y}(T)\| = O(h_1^{p+k})$ (cmp. Thm. 2.4.8)

$(*)$: Thm. 2.4.11 ➤ Extrapolation based on polynomial of the form

$$p(t) = \alpha_0 + \alpha_p h^p + \alpha_{p+1} h^{p+1} + \cdots + \alpha_{p+k-1} h^{p+k-1} \ !$$

- Thm. 2.4.11 requires "‘sufficiently small” h
- Not only $\mathbf{y}(T)$ is of interest but also (approximated) solution $t \mapsto \mathbf{y}(t)$

2.4.4 Local single step extrapolation method

► Application of the extrapolation idea to intervals of a time grid $\mathcal{G} := \{t_0 < t_1 < \cdots < t_N = T\}$ $\leftrightarrow$ macro steps: on $[t_j, t_{j+1}]$, macro step size $H_j := t_{j+1} - t_j$

- Fix sequence $(n_l)_{l=1}^{k+1}$, $n_l \in \mathbb{N}$, e.g., $(1,2,3,4,5,6,\dots)$ $\leftrightarrow$ number of micro steps
- n_l steps of an SSM with initial value $\mathbf{y}_j$, step size $h = \frac{t_{j+1}-t_j}{n_l}$ ➥ $\mathbf{y}_{j+1}^l$, $l = 1,\dots,k+1$
  (SSM = base method, order p)
- Polynomial extrapolation $(*)$ from $(n_l^{-1}, \mathbf{y}_{j+1}^l)$ ➥ $\mathbf{y}_{j+1}$ cmp. Rem. 2.4.4

single step extrapolation method of order $p+k$

MATLAB-CODE : single step, local extrapolation SSM, scalar ODE

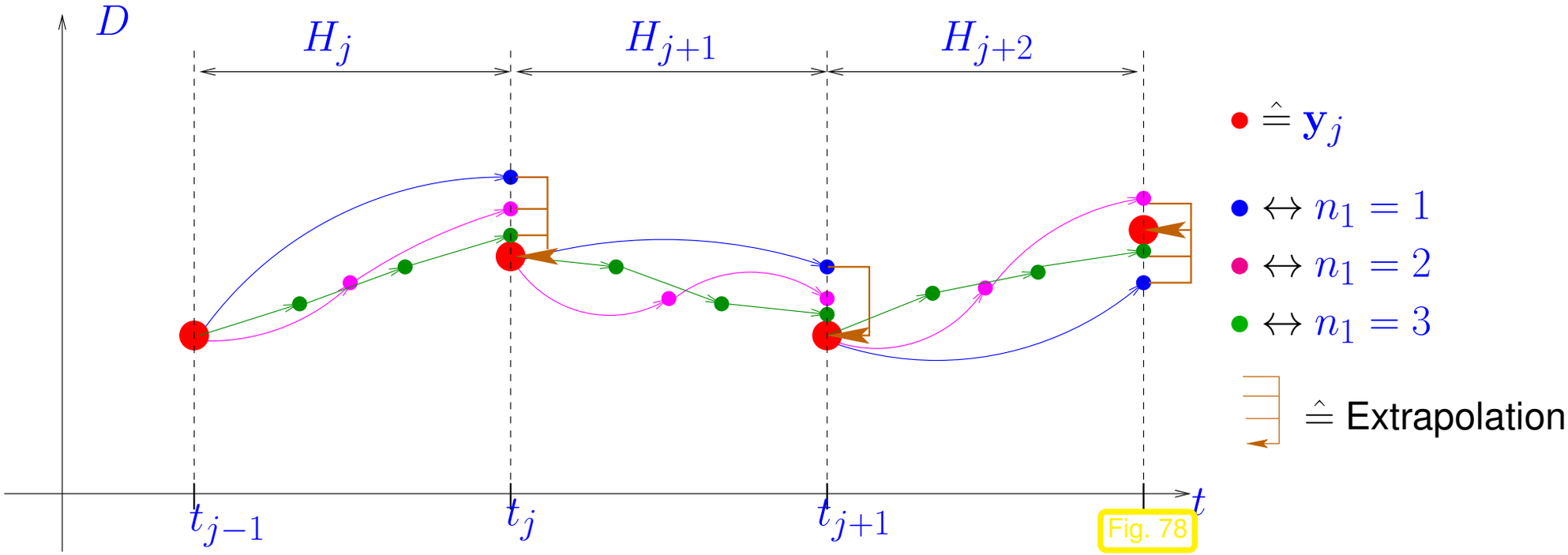
```
function y = expesvstep(esvstep,y,t,h,n)
for i=1:length(n)
  yt(i) = y;
  ht = h/n(i); tt = t;
  for j=1:n(i)
  yt(i) = esvstep(yt(i),tt,ht);
  tt = tt + ht;
end
T = anexpol(yt,1./n,p);
return(T(1));
```

`esvstep(y,t,h)` $\hat{=}$ one step of the base method, step size h, starting from state $(t,\mathbf{y})$:
$\mathtt{esvstep}(\mathbf{y},\mathbf{t},\mathbf{h}) := \mathbf{\Psi}^{t,t+h}\mathbf{y}$

`n` $\hat{=}$ Vector $(n_l)_{l=1}^{k+1}$

`anexpol` $\hat{=}$ more general version of the extrapolation polynomial $p(t) = \alpha_0 + \alpha_p h^p + \alpha_{p+1}h^{p+1} + \dots + \alpha_{p+k-1}h^{p+k-1}$

Process: Local single step extrapolation method $(n_1 = 1, n_2 = 2, n_3 = 3)$



Example 2.4.17 (Extrapolated Euler method).

P. Grohs

rev 63606, February 17, 2014

IVP for logistic differential equation (→ Ex. 1.2.1)

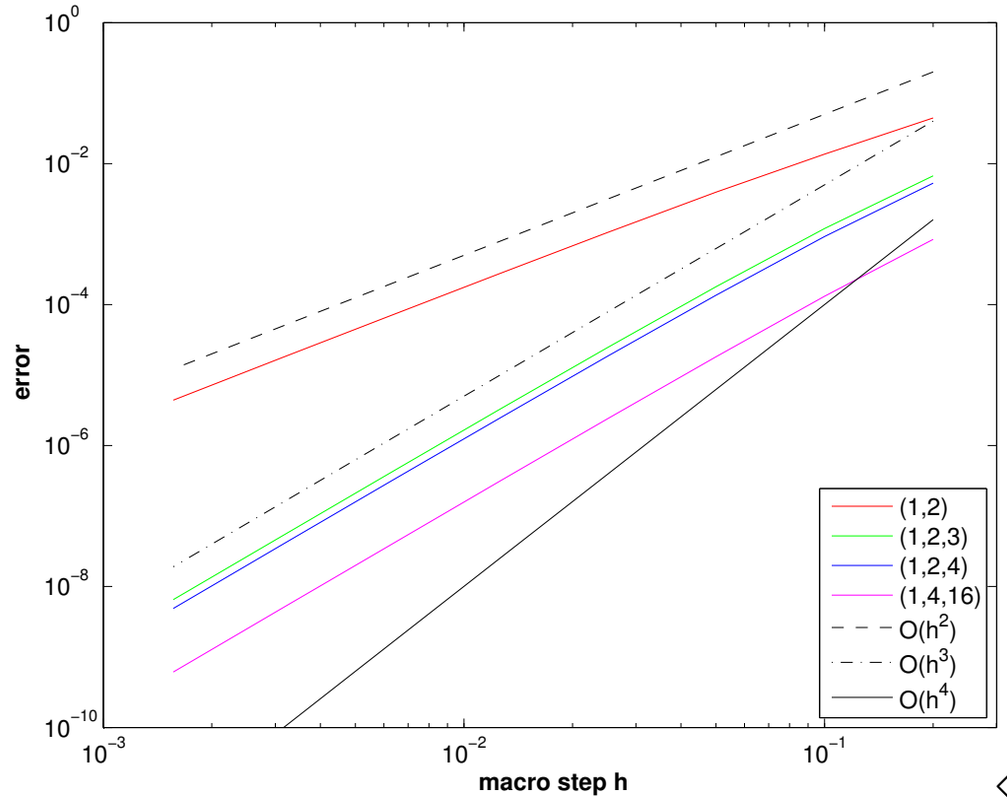
$$\dot{y} = 5y(1-y) \;, \quad y(0) = 0.02 \;.$$

Final time: $T = 1$ Base SSM: explicit Euler method (1.4.2)

Extrapolation: different k, n_l, $l = 1, \dots, k+1$, uniform macro step size h

$$\text{error} \quad \max_j |y(t_j) - y_j| \;.$$

Algebraic convergence of order $k+1$ ▷



◇

P. Grohs

rev 63606, February 17, 2014

Theoretical analysis ↔ verification of the prerequisites of Thm. 2.1.19

Considerations for special case $p = 1$ ↔ Euler method, cmp. Ex. 2.4.17

Notation:

- $t \mapsto \mathbf{y}(t) \mathrel{\hat{=}}$ exact solution in $(t, \mathbf{y}) \in \Omega$

- $H > 0 \hat{=}$ step size of macro steps
- $n_1, \dots, n_{k+1} \hat{=}$ number of micro steps in $[t, t+H]$
- $\mathbf{y}_N \hat{=}$ result of applying N steps of the base single step method to $[t, t+H]$ with uniform step size $h := H/N$ and initial value $\mathbf{y}$
- $\widehat{\mathbf{y}} \hat{=}$ approximate value for $\mathbf{y}(t+H)$ gained by extrapolation from $\mathbf{y}_{n_1}, \mathbf{y}_{n_2}, \dots, \mathbf{y}_{n_{k+1}}$

▶ Consistency error (→ Def. 2.1.11): $\boldsymbol{\tau}(t, \mathbf{y}, H) = \mathbf{y}(t+H) - \widehat{\mathbf{y}}$.

To show: order of consistency $k+1 \quad \leftrightarrow \quad \|\boldsymbol{\tau}(t, \mathbf{y}, H)\| = O(H^{k+2})$

① Local application of Thm. 2.4.11 with $t_0 = t$, $T = t + H$: for sufficiently large $K \in \mathbb{N}$

$$\Rightarrow \quad \mathbf{y}_N - \mathbf{y}(t+H) = \sum_{l=1}^{K} \mathbf{e}_l(t+H) h^l + \mathbf{r}_K(t+H, h) h^{K+1} \ , \quad h = H/N \ ,$$

where ☞ $\|\mathbf{r}_K(t+H, h)\| \leq CH$

☞ $\|\mathbf{e}(t+H)\| \leq CH$

with $C > 0$ independent of t and (sufficiently small) h.

② Therefore, from Thm. 2.4.8

$$\|\widehat{\mathbf{y}} - \mathbf{y}(t+H)\| \leq \|\mathbf{e}_{k+1}(t+H)\| \, h_1 \cdot \dots \cdot h_{k+1} + C \sum_{j=1}^{k} \|\mathbf{r}_j(t+H, h_j)\| \, h_j^{k+2} \ ,$$

P. Grohs

rev 63606, February 17, 2014

where $C > 0$ only depends on the ratios $n_j : n_l$.

$$\Rightarrow \qquad \|\widehat{\mathbf{y}} - \mathbf{y}(t+H)\| \leq CH^{k+2} \,,$$

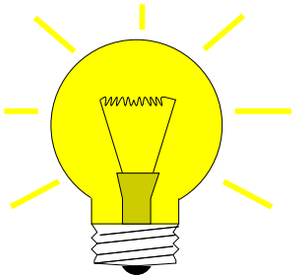
with $C > 0$ independent of H.

Remark 2.4.18 (Extrapolation method as Runge-Kutta method).

Base method: explicit Euler method (1.4.2), order $p = 1$

➜ Polynomial extrapolation for sequence $(1, 2, 3, 4, \ldots, k)$ results in explicit ($\rightarrow$ Def. 2.1.5) Runge-Kutta method ($\rightarrow$ Def. 2.3.5) of order k with $s = k(k-1)/2 + 1$ stages. $\triangle$

2.4.5 Order control



For single step extrapolation method: $k = k(j)$ easy to realize

Idea: [(local) order control]

> "Increase locally the order until it does not pay off anymore $(*)$"

$(*)$ Heuristic criterion (based on Aitken-Neville extrapolation tableau (2.4.6)):

$$\left\| T_{k,k-1} - T_{k,k} \right\| \leq \texttt{TOL} \cdot \left\| T_{k,k} \right\| \quad \text{for tolerance } \texttt{TOL} > 0 \,.$$

second best approximation — best approximation

MATLAB-CODE : Adaptive Euler extrapolation method

```
function [y,k] = eulexstep(f,y,H,TOL)

kmax = 1000;
T{1} = y + H*f(y);
for i=2:kmax
  T{i} = y; h = H/i;
  for k=1:i, T{i} = T{i} + h*f(T{i}); end
  for l=i-1:-1:1
    T{l} = T{l+1} + (T{l+1}-T{l})/(i/l-1);
  end
  if (norm(T{1}-T{2}) < TOL*norm(T{1}))
  y = T{1}; k = i; return;
end
```

Adaptive

Euler extrapolation method:

(for autonomous IVP)

macro steps of size H

Input arguments:

- f : function handle f=@(y) for right hand side
- y : initial value for $t = 0$
- TOL : tolerance

Return values:

- y : approximation for $t = H$
- k : utilized extrapolation depth

P. Grohs

rev 63606, February 17, 2014

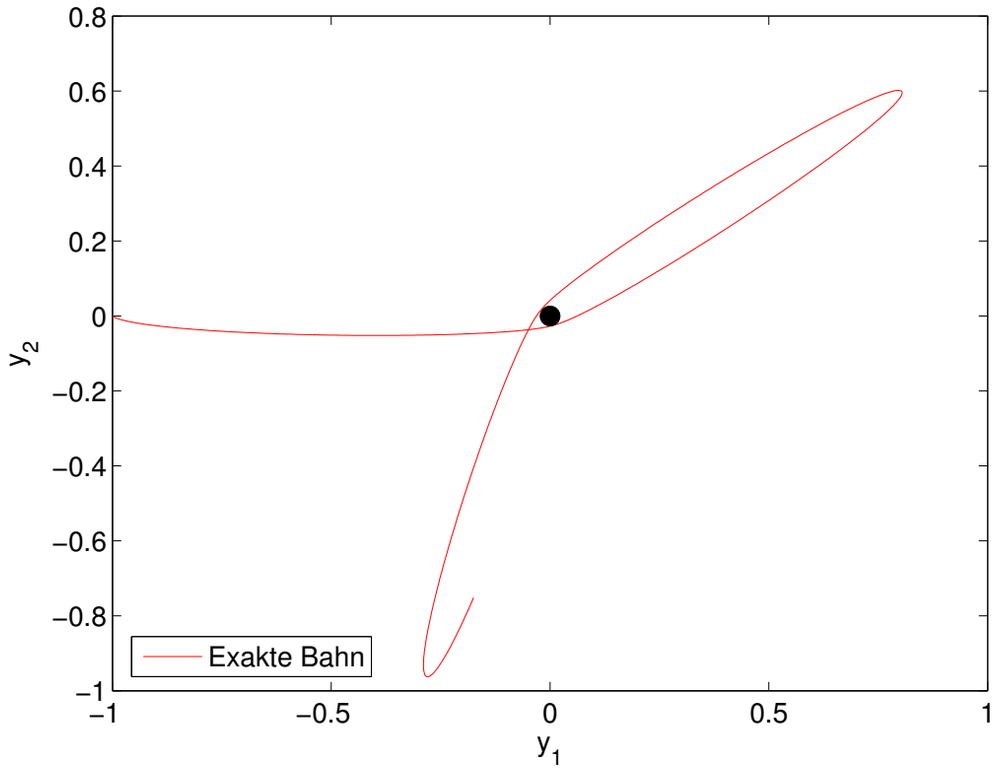
Note: Simple extension of the extrapolation tableau by an additional row

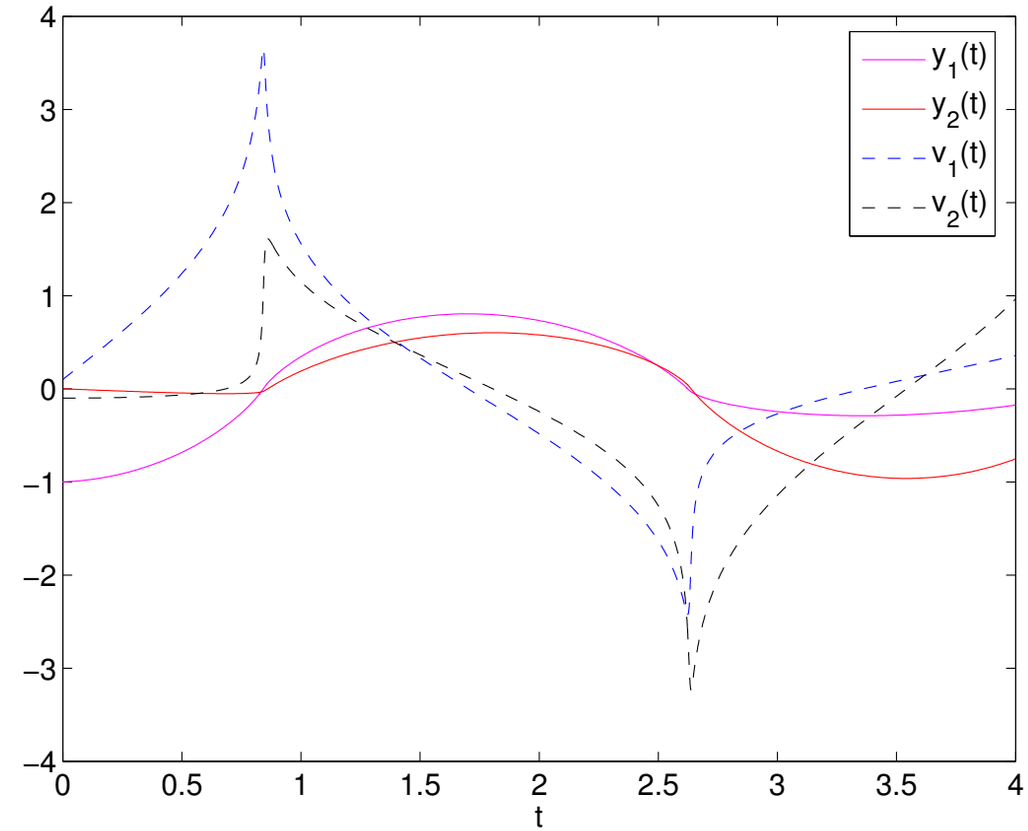
Example 2.4.19 (Euler extrapolation method with order control).

Movement of a charged particle in the field of a straight wire = line charge (conservative central field, center $\binom{0}{0}$, potential $U(\boldsymbol{x}) := -2\log\|\boldsymbol{x}\|$): $\rightarrow$ Ex. 1.2.25

$$\ddot{\mathbf{y}} = -\frac{2\mathbf{y}}{\|\mathbf{y}\|^2} \quad \Rightarrow \quad \dot{\begin{pmatrix}\mathbf{y}\\ \mathbf{v}\end{pmatrix}} = \begin{pmatrix}\mathbf{v}\\ -\frac{2\mathbf{y}}{\|\mathbf{y}\|^2}\end{pmatrix} \quad , \quad \mathbf{y}(0) = \begin{pmatrix}-1\\ 0\end{pmatrix}, \mathbf{v}(0) = \begin{pmatrix}0.1\\ -0.1\end{pmatrix} .$$

Initial value: $\mathbf{y}(0) = (-1, 0, 0.1, -0.1)$, final time: $T = 4$





Observation: "peaks" in the solution component $\mathbf{v}(t)$ (= time-local characteristic)

Order-adaptive Euler extrapolation method ($\texttt{TOL} = 0.01$), uniform macro step size $H = 0.02$

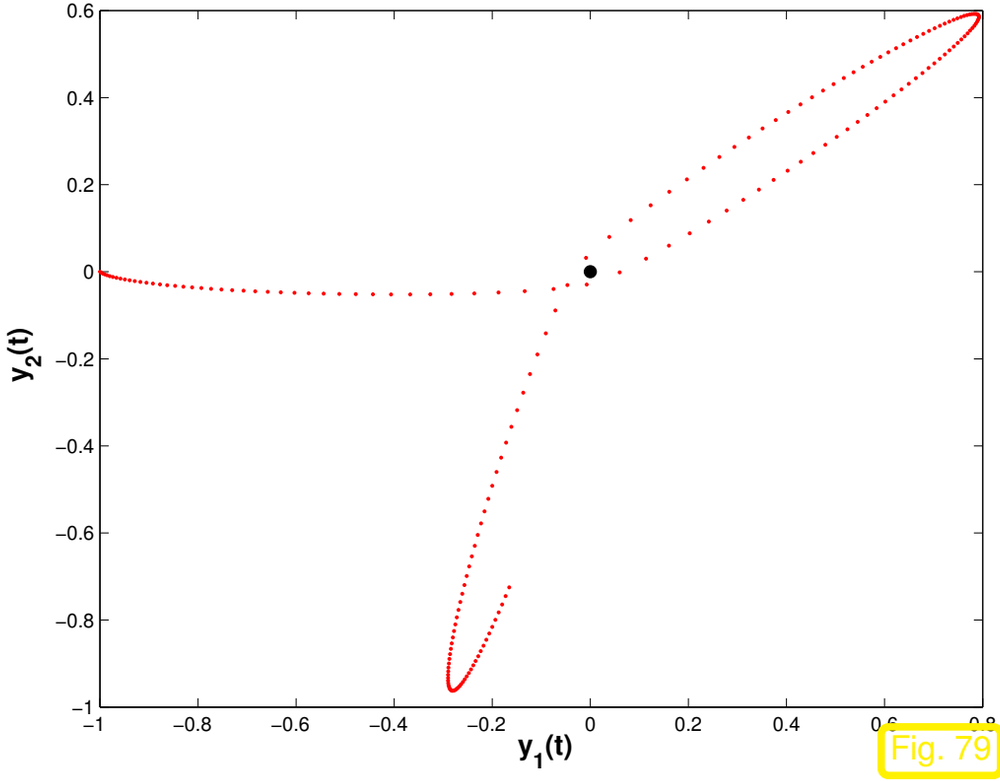


Fig. 79

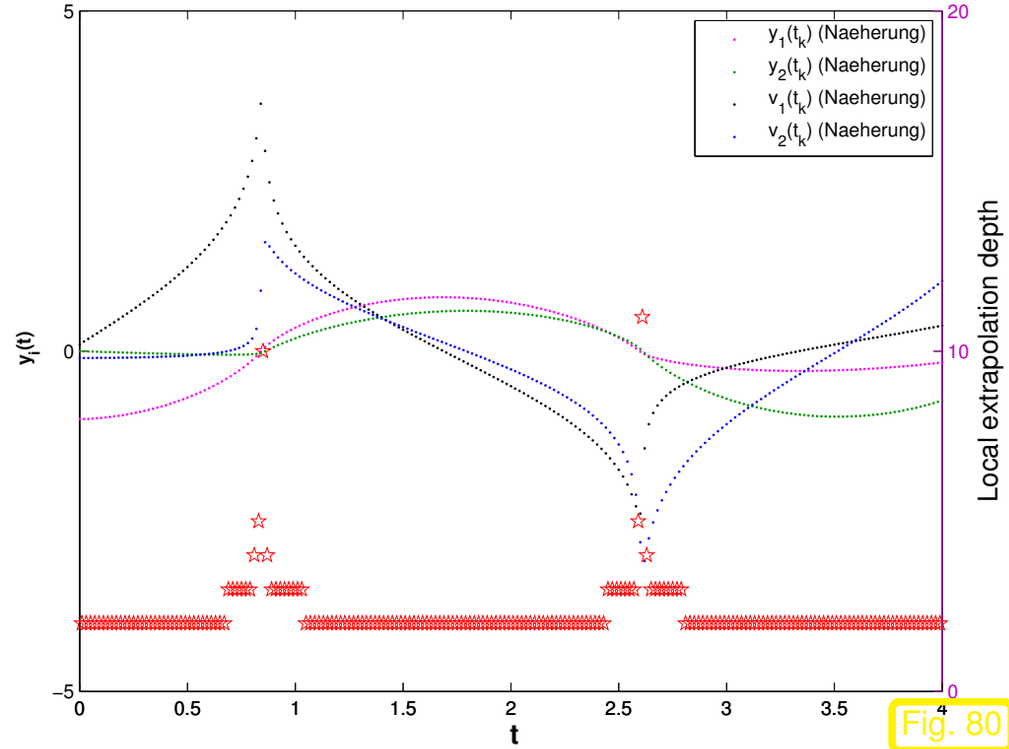


Fig. 80

▷ Automatic increase in order at “critical points”

◇

2.4.6 Extrapolation of reversible single step methods

Example 2.4.20 (Extrapolated implicit midpoint rule).

- Initial value problem from Ex. 1.4.9 (logistic differential equation, compare Ex. 1.2.1), $\lambda = 10$, $y_0 = 0.01$, from $[0,1]$ ($T = 1$)
- Single step method: implicit midpoint rule (1.4.19)
- Global extrapolation ($\rightarrow$ section 2.4.3) of $y_h(T)$ from solutions obtained by uniform step size h/n_i

  Note: Extrapolation based on the standard tableau (2.4.5)

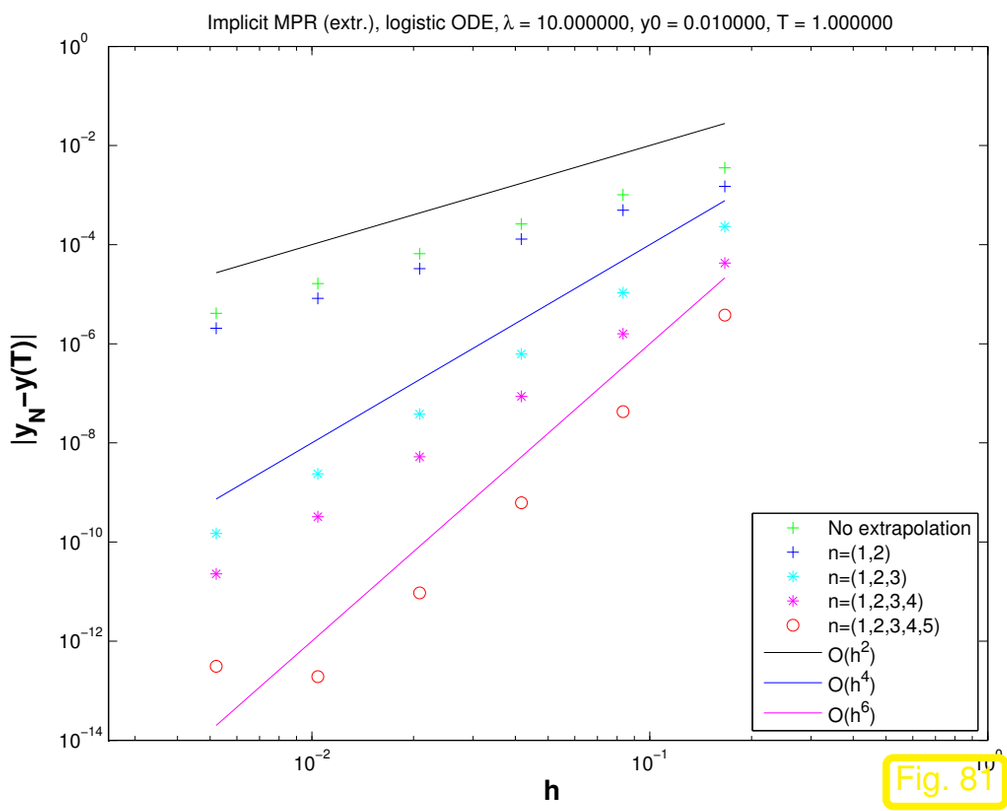


Fig. 81

◁ Increase in order only in every second extrapolation step

Increase in order by two, respectively

◇

It is easy to explain the observation from Ex. 2.4.20, if

$$y_h(T) = y(T) + \alpha_1 h^2 + \alpha_2 h^4 + \alpha_6 h^6 + \cdots \ .$$

Example 2.4.21 (Global h^2 extrapolation for implicit midpoint rule).

- (Almost) like Ex. 2.4.20
- NEW:

y_N from extrapolation in h^2

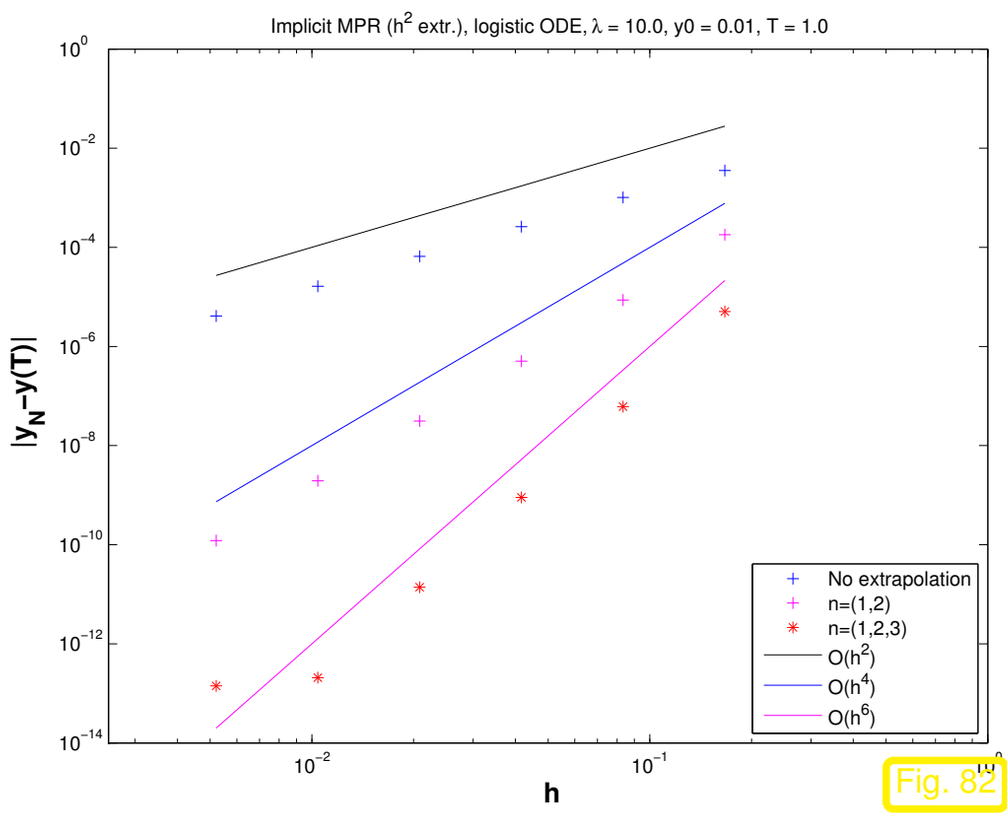


Fig. 82

◁ Increase in order by **two** in every extrapolation step **!**

◇

Theorem 2.4.22 (Asymptotic expansion of the discretization error in h^2)**.**
Let $\mathbf{y}_h(t)$, $t \in$ equidistant time grid with step size $h > 0$ on $[t_0, T]$, denote an approximate solution to the initial value problem $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$, $\mathbf{y}(t_0) = \mathbf{y}_0$, with exact solution $t \mapsto \mathbf{y}(t)$, that was generated by a reversible *single step method ($\rightarrow$ Def. 2.1.27).*
Then there exist a $K \in \mathbb{N}$ (dependent on the smoothness of $\mathbf{f}$) and smooth functions $\mathbf{e}_i : J(t_0, \mathbf{y}_0) \mapsto \mathbb{R}^d$, $i = 1, \dots, K$, with $\mathbf{e}_i(0) = 0$ and (for sufficiently small h) uniformly bounded functions $(T, h) \mapsto \mathbf{r}_k(T, h)$, $0 \leq k \leq K$ such that

$$\mathbf{y}_h(T) - \mathbf{y}(T) = \sum_{l=1}^{k} \mathbf{e}_l(T) h^{2l} + \mathbf{r}_k(T, h) h^{2k+2} \quad \text{for small } h \ .$$

Hereby, there holds $\quad \|\mathbf{r}_k(T, h)\| = O(T - t_0) \quad$ *for $T - t_0 \to 0$ uniformly in $h < T$.*

Proof. Compare [12, Satz 4.42] □

P. Grohs

rev 63606, February 17, 2014

Remark 2.4.23 (DIFEX).

Practical extrapolation methods are based on *explicit* methods, for which the corresponding error has an asymptotic expansion in h^2 (a special trapezoidal rule)

➢ DIFEX algorithm [12, Sect. 4.3.3]

△

2.5 Splitting methods [21, Sect. 2.5]

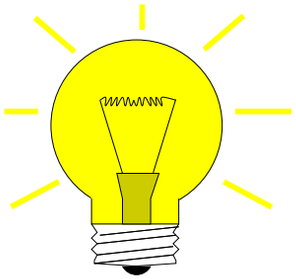
Autonomous IVP with additively decomposed right hand side:

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) + \mathbf{g}(\mathbf{y}) \quad , \quad \mathbf{y}(0) = \mathbf{y}_0 \; , \tag{2.5.1}$$

with $\mathbf{f} : D \subset \mathbb{R}^d \mapsto \mathbb{R}^d$, $\mathbf{g} : D \subset \mathbb{R}^d \mapsto \mathbb{R}^d$ "sufficiently smooth", locally Lipschitz continuous (→ Def. 1.3.2)

$$\text{(Continuous) evolutions:} \quad \begin{array}{lll} \mathbf{\Phi}_f^t & \leftrightarrow & \text{diff. eqn.} \quad \dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) \; , \\ \mathbf{\Phi}_g^t & \leftrightarrow & \text{diff. eqn.} \quad \dot{\mathbf{y}} = \mathbf{g}(\mathbf{y}) \; . \end{array}$$

Assumption: $\mathbf{\Phi}_f^t$, $\mathbf{\Phi}_g^t$ (analytically) known



Idea: construct single step methods with discrete evolutions

Lie-Trotter splitting: $$\mathbf{\Psi}^h = \mathbf{\Phi}_g^h \circ \mathbf{\Phi}_f^h \;, \tag{2.5.2}$$

Strang splitting: $$\mathbf{\Psi}^h = \mathbf{\Phi}_f^{h/2} \circ \mathbf{\Phi}_g^h \circ \mathbf{\Phi}_f^{h/2} \;. \tag{2.5.3}$$

(2.5.2) $\leftrightarrow$

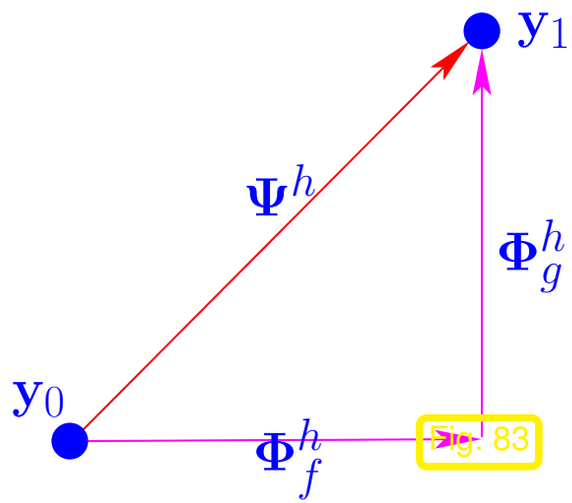


Fig. 83

(2.5.3) $\leftrightarrow$

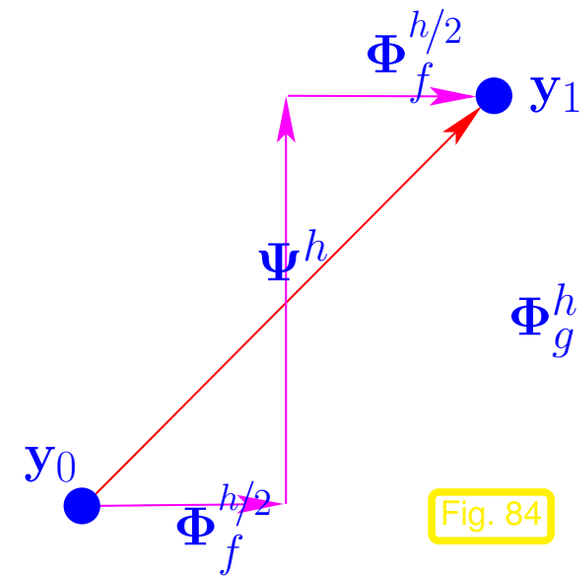


Fig. 84

Example 2.5.4 (Convergence of simple splitting methods).

$$\dot{y} = \underbrace{\lambda y(1-y)}_{=:f(y)} + \underbrace{\sqrt{1-y^2}}_{=:g(y)} \quad , \quad y(0) = 0 \;.$$

- $\Phi_f^t y = \dfrac{1}{1+(y^{-1}-1)e^{-\lambda t}}$, $t>0, y\in]0,1]$ (logistic differential equation (2.2.57))
- $\Phi_g^t y = \begin{cases} \sin(t+\arcsin(y)) & \text{, if } t+\arcsin(y) < \frac{\pi}{2}\,, \\ 1 & \text{, otherwise,} \end{cases}$ $t>0, y\in[0,1]$.

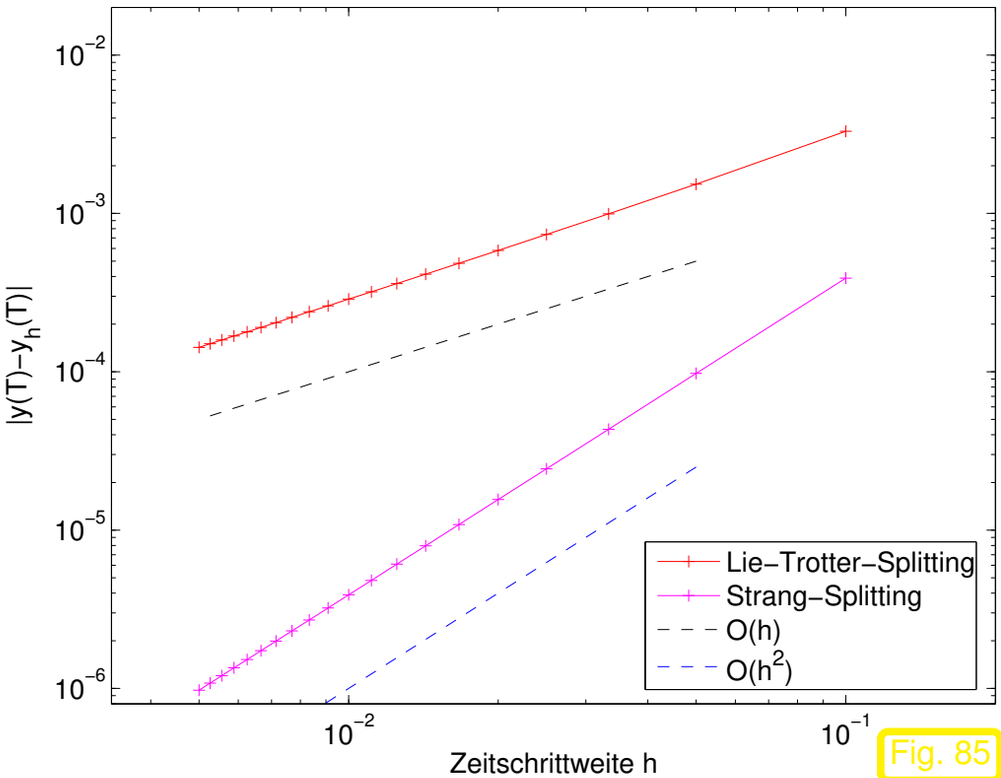


Fig. 85

Numerical experiment:

$T=1$, $\lambda=1$, comparison of splitting methods (constant step size) with highly accurate numerical solution obtained by

```
f=@(t,x) λ*x*(1-x)+sqrt(1-x^2);
options=odeset('reltol',1.0e-10,...
               'abstol',1.0e-12);
[t,yex]=ode45(f,[0,1],y0,options);
```

◁ Behavior of error at final time $T=1$ ◇

Theorem 2.5.5 (Order of consistency of simple splitting methods)**.** *The SSM* (2.5.2) *and* (2.5.3) *have the order of consistency (→ Def. 2.1.13) 1 and 2, respectively.*

Proof. For the consistency error ($\rightarrow$ Def. 2.1.11), we have to show according to Def. 2.1.13 (we consider autonomous ODEs!)

$$\|\tau(t,\mathbf{y},h)\| = \left\|\mathbf{\Phi}^h\mathbf{y} - \mathbf{\Psi}^h\mathbf{y}\right\| = \begin{cases} O(h^2) & \text{for } \mathbf{\Psi} \text{ from (2.5.2)} , \\ O(h^3) & \text{for } \mathbf{\Psi} \text{ from (2.5.3)} . \end{cases}$$

The proof here is done for Strang splitting.

Common assumption: $\mathbf{f}$, $\mathbf{g}$ sufficiently smooth.

Technique: **Taylor expansion**

Taylor expansion of the exact evolution w.r.t. h, cmp. (2.3.25):

$$\begin{aligned} \mathbf{\Phi}^h\mathbf{y} =& \mathbf{y} + \dot{\mathbf{y}}(0)h + \tfrac{1}{2}\ddot{\mathbf{y}}(0)h^2 + O(h^3) \\ =& \mathbf{y} + h(\mathbf{f}(\mathbf{y}) + \mathbf{g}(\mathbf{y})) + \tfrac{1}{2}h^2(D\mathbf{f}(\mathbf{y}) + D\mathbf{g}(\mathbf{y}))(\mathbf{f}(\mathbf{y}) + \mathbf{g}(\mathbf{y})) + O(h^3) . \end{aligned} \tag{2.5.6}$$

Taylor expansion of the partial evolutions $\mathbf{\Phi}_f^h$, $\mathbf{\Phi}_g^h$ w.r.t. h, cmp. (2.3.25)

$$\mathbf{\Phi}_f^h\mathbf{y} = \mathbf{y} + h\mathbf{f}(\mathbf{y}) + \tfrac{1}{2}h^2 D\mathbf{f}(\mathbf{y})\mathbf{f}(\mathbf{y}) + O(h^3) , \tag{2.5.7}$$

$$\mathbf{\Phi}_g^h\mathbf{y} = \mathbf{y} + h\mathbf{g}(\mathbf{y}) + \tfrac{1}{2}h^2 D\mathbf{g}(\mathbf{y})\mathbf{g}(\mathbf{y}) + O(h^3) . \tag{2.5.8}$$

Successive substitution of Taylor expansions, with multiplicative factors h^k allowing for an early truncation of the substituted Taylor expansion, compare the Taylor expansion of the Runge-Kutta incre-

ments in Ex. 2.3.24, (2.3.27).

$$
\begin{aligned}
\mathbf{\Psi}^h\mathbf{y} =& \mathbf{\Phi}_f^{h/2}(\mathbf{\Phi}_g^h(\mathbf{\Phi}_f^{h/2}\mathbf{y}))\\
\overset{❶}{=}& \mathbf{\Phi}_f^{h/2}(\mathbf{\Phi}_g^h(\mathbf{y}+h/2\mathbf{f}(\mathbf{y})+\tfrac{1}{8}h^2D\mathbf{f}(\mathbf{y})\mathbf{f}(\mathbf{y})+O(h^3)))\\
\overset{❷}{=}& \mathbf{\Phi}_f^{h/2}\Big(\mathbf{y}+h/2\mathbf{f}(\mathbf{y})+\tfrac{1}{8}h^2D\mathbf{f}(\mathbf{y})\mathbf{f}(\mathbf{y})+O(h^3)+h\mathbf{g}(\mathbf{y}+h/2\mathbf{f}(\mathbf{y})+O(h^2))\\
&+\tfrac{1}{2}h^2D\mathbf{g}(\mathbf{y}+O(h))\mathbf{g}(\mathbf{y}+O(h))+O(h^3)\Big)\\
\overset{❸}{=}& \mathbf{\Phi}_f^{h/2}\Big(\mathbf{y}+h/2\mathbf{f}(\mathbf{y})+h\mathbf{g}(\mathbf{y})+\tfrac{1}{8}h^2D\mathbf{f}(\mathbf{y})\mathbf{f}(\mathbf{y})+\tfrac{1}{2}h^2D\mathbf{g}(\mathbf{y})\mathbf{f}(\mathbf{y})+\\
&\tfrac{1}{2}h^2D\mathbf{g}(\mathbf{y})\mathbf{g}(\mathbf{y})+O(h^3)\Big)\\
\overset{❹}{=}& \mathbf{y}+\tfrac{h}{2}\mathbf{f}(\mathbf{y})+h\mathbf{g}(\mathbf{y})+\tfrac{1}{8}h^2D\mathbf{f}(\mathbf{y})\mathbf{f}(\mathbf{y})+\tfrac{1}{2}h^2D\mathbf{g}(\mathbf{y})\mathbf{f}(\mathbf{y})+\tfrac{1}{2}h^2D\mathbf{g}(\mathbf{y})\mathbf{g}(\mathbf{y})+O(h^3)+\\
&h/2\mathbf{f}(\mathbf{y}+\tfrac{h}{2}\mathbf{f}(\mathbf{y})+h\mathbf{g}(\mathbf{y}))+O(h^2))+\tfrac{1}{8}h^2D\mathbf{f}(\mathbf{y}+O(h))\mathbf{f}(\mathbf{y}+O(h))\\
\overset{❺}{=}& \mathbf{y}+\tfrac{h}{2}\mathbf{f}(\mathbf{y})+h\mathbf{g}(\mathbf{y})+\tfrac{1}{8}h^2D\mathbf{f}(\mathbf{y})\mathbf{f}(\mathbf{y})+\tfrac{1}{2}h^2D\mathbf{g}(\mathbf{y})\mathbf{f}(\mathbf{y})+\tfrac{1}{2}h^2D\mathbf{g}(\mathbf{y})\mathbf{g}(\mathbf{y})\\
&+\tfrac{1}{2}h\mathbf{f}(\mathbf{y})+\tfrac{1}{4}h^2D\mathbf{f}(\mathbf{y})\mathbf{f}(\mathbf{y})+\tfrac{1}{2}h^2D\mathbf{f}(\mathbf{y})\mathbf{g}(\mathbf{y})+\tfrac{1}{8}h^2D\mathbf{f}(\mathbf{y})\mathbf{f}(\mathbf{y})+O(h^3)\\
=& \mathbf{y}+h(\mathbf{f}(\mathbf{y})+\mathbf{g}(\mathbf{y}))+\tfrac{1}{2}h^2\big(D\mathbf{f}(\mathbf{y})\mathbf{f}(\mathbf{y})+D\mathbf{g}(\mathbf{y})\mathbf{f}(\mathbf{y})+D\mathbf{f}(\mathbf{y})\mathbf{g}(\mathbf{y})+D\mathbf{g}(\mathbf{y})\mathbf{g}(\mathbf{y})\big)\\
&+O(h^3)\,.
\end{aligned}
$$

❶ Apply (2.5.7) to $\mathbf{\Phi}_f^{h/2}\mathbf{y}$.

❷ Use (2.5.8) with $\mathbf{y} \leftarrow \mathbf{y} + h/2\mathbf{f}(\mathbf{y}) + \frac{1}{8}h^2 D\mathbf{f}(\mathbf{y})\mathbf{f}(\mathbf{y}) + O(h^3)$, to expand $(\mathbf{\Phi}_g^h(\mathbf{y} + h/2\mathbf{f}(\mathbf{y}) + \frac{1}{8}h^2 D\mathbf{f}(\mathbf{y})\mathbf{f}(\mathbf{y}) + O(h^3))$. Neglect terms $O(h^3)$.

❸ Taylor expansion (in h) of $\mathbf{g}$ and $D\mathbf{g}$ around $\mathbf{y}$.

❹ Use (2.5.7) with $\mathbf{y} \leftarrow \mathbf{y} + h/2\mathbf{f}(\mathbf{y}) + h\mathbf{g}(\mathbf{y}) + \frac{1}{8}h^2 D\mathbf{f}(\mathbf{y})\mathbf{f}(\mathbf{y}) + \frac{1}{2}h^2 D\mathbf{g}(\mathbf{y})\mathbf{f}(\mathbf{y}) + \frac{1}{2}h^2 D\mathbf{g}(\mathbf{y})\mathbf{g}(\mathbf{y}) + O(h^3)$ and neglect terms in $O(h^3)$.

❺ Taylor expansion (in h) of $\mathbf{f}$ and $D\mathbf{f}$ around $\mathbf{y}$.

Comparison with (2.5.6) implies the claim for Strang splitting. □

Remark 2.5.9 (Reversible single step Strang splitting method).

△

Example 2.5.10 (Splitting method for mechanical systems).

$$\text{Newton's equations of motion} \quad \ddot{\mathbf{r}} = a(\mathbf{r}) \overset{(1.1.10)}{\Longleftrightarrow} \dot{\mathbf{y}} := \dot{\begin{pmatrix} \mathbf{r} \\ \mathbf{v} \end{pmatrix}} = \begin{pmatrix} \mathbf{v} \\ a(\mathbf{r}) \end{pmatrix} =: \mathbf{F}(\mathbf{y}) \ .$$

Splitting: $$F(\mathbf{y}) = \underbrace{\begin{pmatrix} 0 \\ a(\mathbf{r}) \end{pmatrix}}_{=:\mathbf{f}(\mathbf{y})} + \underbrace{\begin{pmatrix} \mathbf{v} \\ 0 \end{pmatrix}}_{=:\mathbf{g}(\mathbf{y})} .$$

► $$\mathbf{\Phi}_f^t \begin{pmatrix} \mathbf{r}_0 \\ \mathbf{v}_0 \end{pmatrix} = \begin{pmatrix} \mathbf{r}_0 \\ \mathbf{v}_0 + ta(\mathbf{r}_0) \end{pmatrix} \quad , \quad \mathbf{\Phi}_g^t \begin{pmatrix} \mathbf{r}_0 \\ \mathbf{v}_0 \end{pmatrix} = \begin{pmatrix} \mathbf{r}_0 + t\mathbf{v}_0 \\ \mathbf{v}_0 \end{pmatrix} .$$

► Lie-Trotter splitting 2.5.2 Symplectic Euler method

$$\mathbf{\Psi}^h \begin{pmatrix} \mathbf{r} \\ \mathbf{v} \end{pmatrix} = \left(\mathbf{\Phi}_g^h \circ \mathbf{\Phi}_f^h \right) \begin{pmatrix} \mathbf{r} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \mathbf{r} + h(\mathbf{v} + ha(\mathbf{r})) \\ \mathbf{v} + ha(\mathbf{r}) \end{pmatrix} . \tag{2.5.11}$$

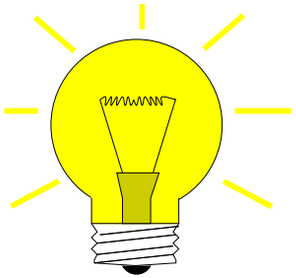
► Strang splitting 2.5.3

$$\mathbf{\Psi}^h \begin{pmatrix} \mathbf{r} \\ \mathbf{v} \end{pmatrix} = \left(\mathbf{\Phi}_g^{h/2} \circ \mathbf{\Phi}_f^h \circ \mathbf{\Phi}_g^{h/2} \right) \begin{pmatrix} \mathbf{r} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \mathbf{r} + h\mathbf{v} + \frac{1}{2}h^2 a(\mathbf{r} + \frac{1}{2}h\mathbf{v}) \\ \mathbf{v} + ha(\mathbf{r} + \frac{1}{2}h\mathbf{v}) \end{pmatrix} . \tag{2.5.12}$$

= Single step formulation of the Störmer-Verlet method (1.4.27), compare Rem. 1.4.33 **!**

$$(2.5.12) \quad \longleftrightarrow \quad \begin{aligned} \mathbf{r}_{k+\frac{1}{2}} &= \mathbf{r}_k + \tfrac{1}{2}h\mathbf{v}_k \ , \\ \mathbf{v}_{k+1} &= \mathbf{v}_k + ha(\mathbf{r}_{k+\frac{1}{2}}) \ , \\ \mathbf{r}_{k+1} &= \mathbf{r}_{k+\frac{1}{2}} + \tfrac{1}{2}h\mathbf{v}_{k+1} \ . \end{aligned} \tag{2.5.13}$$

Objection: Splitting methods are only useful in special cases as it is often not possible to evaluate the exact evolutions $\mathbf{\Phi}_f$ and $\mathbf{\Phi}_g$ analytically.

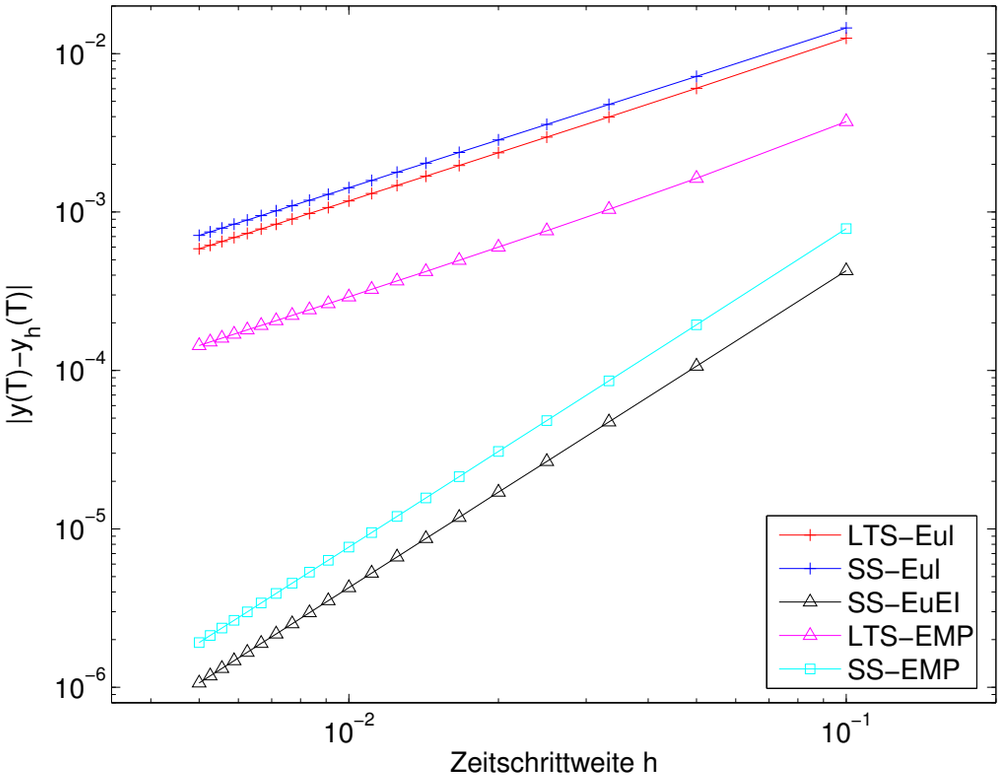


Idea: substitute

$$\begin{array}{ccc} \text{exact evolution} & \longrightarrow & \text{discrete evolution} \\ \mathbf{\Phi}_g^h,\quad \mathbf{\Phi}_f^h & \longrightarrow & \mathbf{\Psi}_g^h,\quad \mathbf{\Psi}_f^h \end{array}$$

Example 2.5.14 (Inexact splitting method). continuation Ex. 2.5.4

IVP of Ex. 2.5.4, inexact splitting method based on different inexact base methods:



LTS-Eul	explicit Euler method (2.2.1) $\rightarrow$ $\mathbf{\Psi}^h_{h,g}$, $\mathbf{\Psi}^h_{h,f}$ + Lie-Trotter splitting (2.5.2)
SS-Eul	explicit Euler method (2.2.1) $\rightarrow$ $\mathbf{\Psi}^h_{h,g}$, $\mathbf{\Psi}^h_{h,f}$ + Strang splitting (2.5.3)
SS-EuEI	Strang splitting (2.5.3): explicit Euler method (2.2.1) $\circ$ exact evolution $\mathbf{\Phi}^h_g$ $\circ$ implicit Euler method (2.2.1)
LTS-EMP	explicit midpoint rule (2.3.4) $\rightarrow$ $\mathbf{\Psi}^h_{h,g}$, $\mathbf{\Psi}^h_{h,f}$ + Lie-Trotter splitting (2.5.2)
SS-EMP	explicit midpoint rule (2.3.4) $\rightarrow$ $\mathbf{\Psi}^h_{h,g}$, $\mathbf{\Psi}^h_{h,f}$ + Strang splitting (2.5.3)

☞ Order of the splitting methods is bounded by order of consistency of $\mathbf{\Phi}^h_f$, $\mathbf{\Phi}^h_g$.

Exception: SS-EuEI: reversible method ➢ order of consistency ≥ 2 according to Thm. 2.1.29 ◇

2.6 Step size control [12, Kap. 5], [24, Sect. 2.8]

Example 2.6.1 (Numerical integration in the case of blow-up).

Scalar autonomous IVP $\rightarrow$ Ex. 1.3.11

$$\dot{y} = y^2 \;, \quad y(0) = y_0 > 0 \;.$$

$$\blacktriangleright \quad y(t) = \frac{y_0}{1 - y_0 t} \;, \quad t < 1/y_0 \;.$$

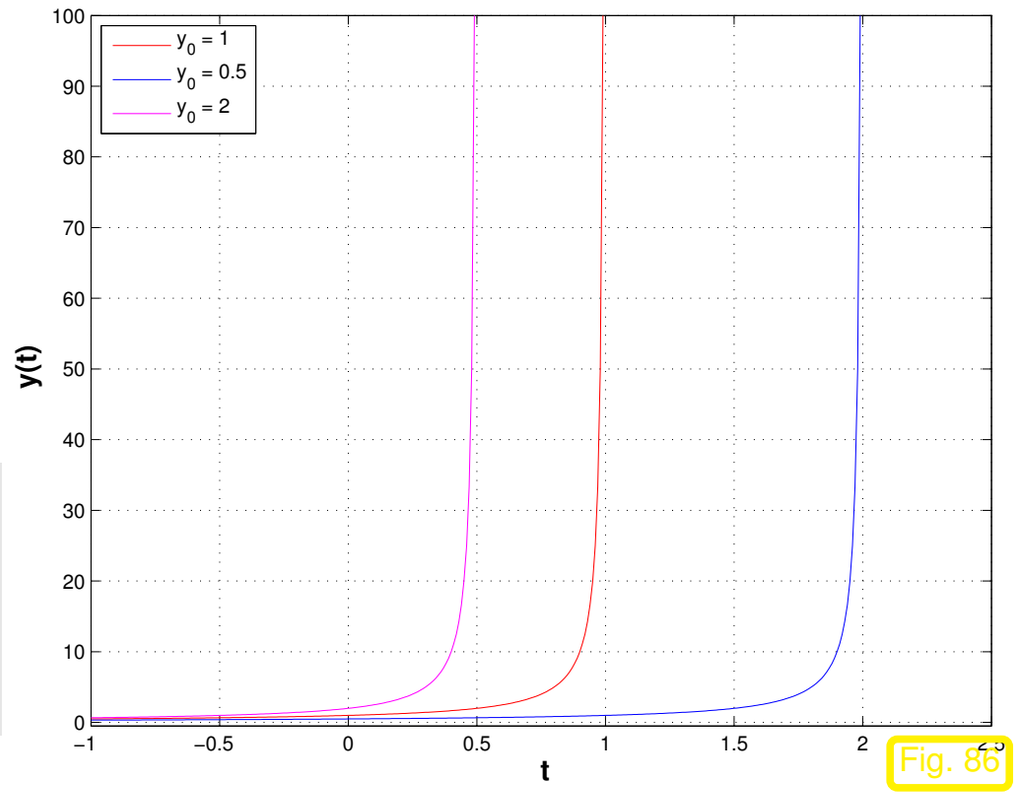


Fig. 86

The solution exists only for finite time and suffers from blow-up, compare Def. 1.3.1: $\lim_{t \to 1/y_0} y(t) = \infty$: $J(y_0) =]-\infty, 1/y_0]$!

Challenge: How should the time grid $\{t_0 < t_1 < \cdots < t_{N-1} < t_N\}$ for a SSM be chosen, if $J(y_0)$ is not known a priori and if it is not clear whether there will be a blow-up?

Thought experiment: how does a single step Runge-Kutta method (→ Def. 2.3.5) behave when applied to above IVP using *uniform (equidistant)* time steps?

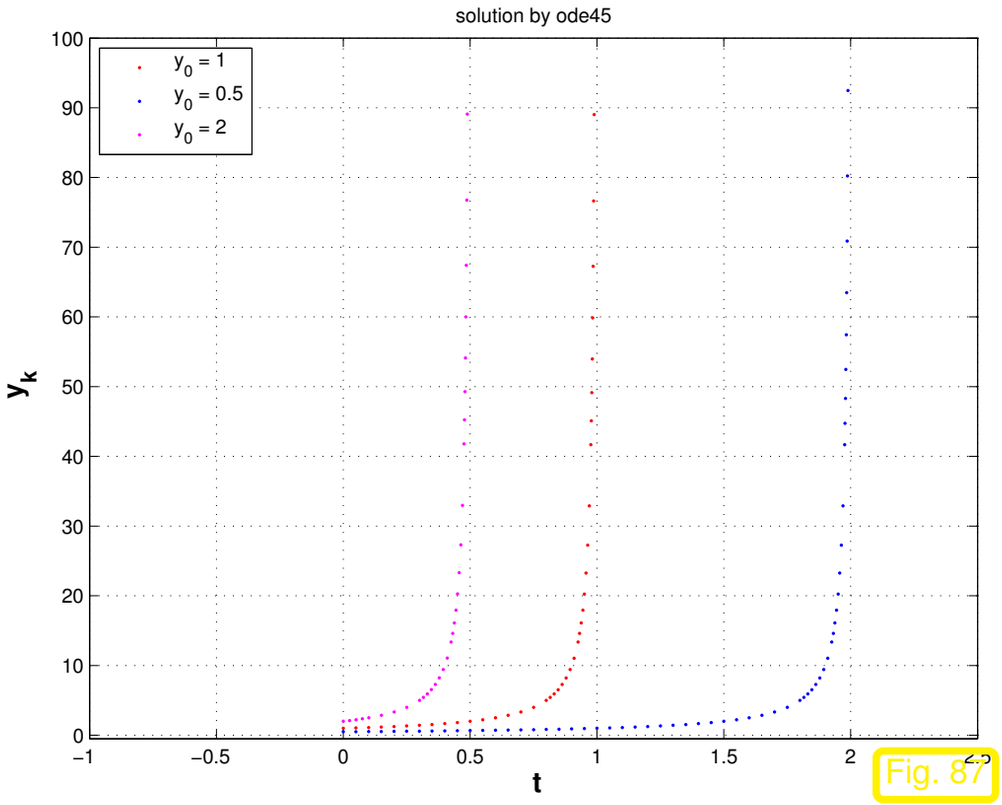


Fig. 87

```
fun = @(t,y) y.^2;
[t1,y1] = ode45(fun,[0 2],1);
[t2,y2] = ode45(fun,[0 2],0.5);
[t3,y3] = ode45(fun,[0 2],2);
```

MATLAB warnings:

```
  Warning: Failure at t=9.999694e-01.  Unable to meet integration
  tolerances without reducing the step size below the smallest
  value allowed (1.776357e-15) at time t.
> In ode45 at 371
  In simpleblowup at 22
```

P. Grohs

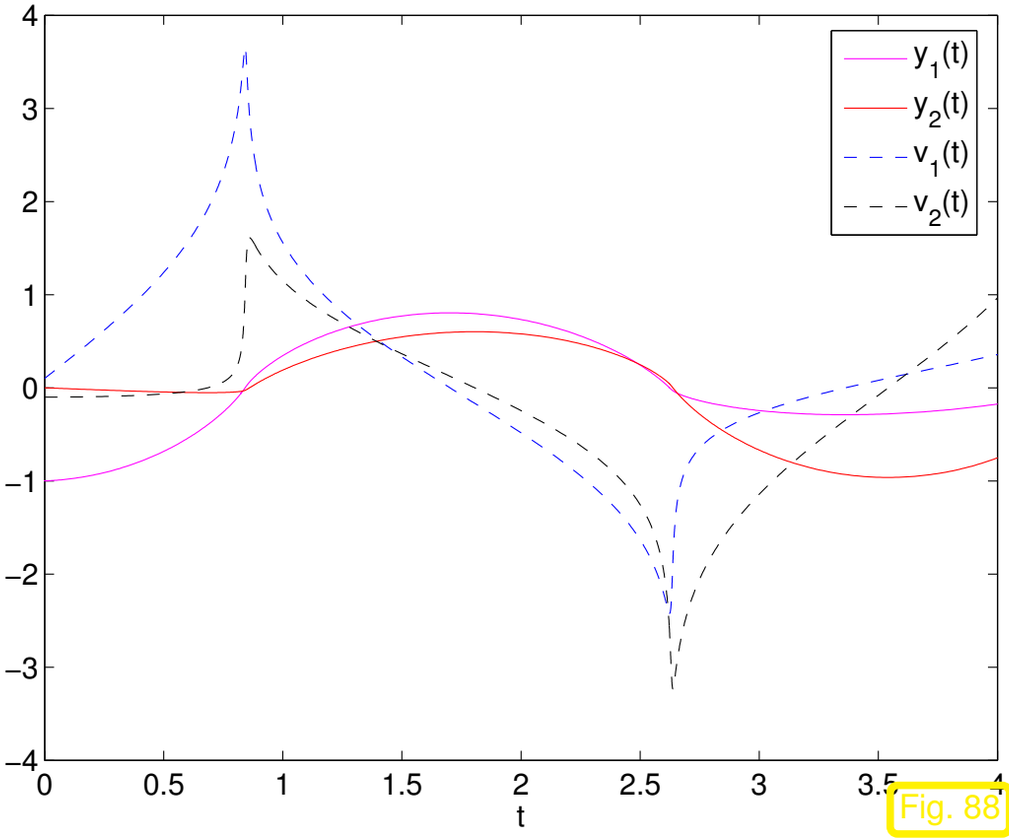
rev 63606, February 17, 2014

```
Warning: Failure at t=1.999970e+00.  Unable to meet integration
  tolerances without reducing the step size below the smallest
  value allowed (3.552714e-15) at time t.
> In ode45 at 371
  In simpleblowup at 23
Warning: Failure at t=4.999660e-01.  Unable to meet integration
  tolerances without reducing the step size below the smallest
  value allowed (8.881784e-16) at time t.
> In ode45 at 371
  In simpleblowup at 24
```

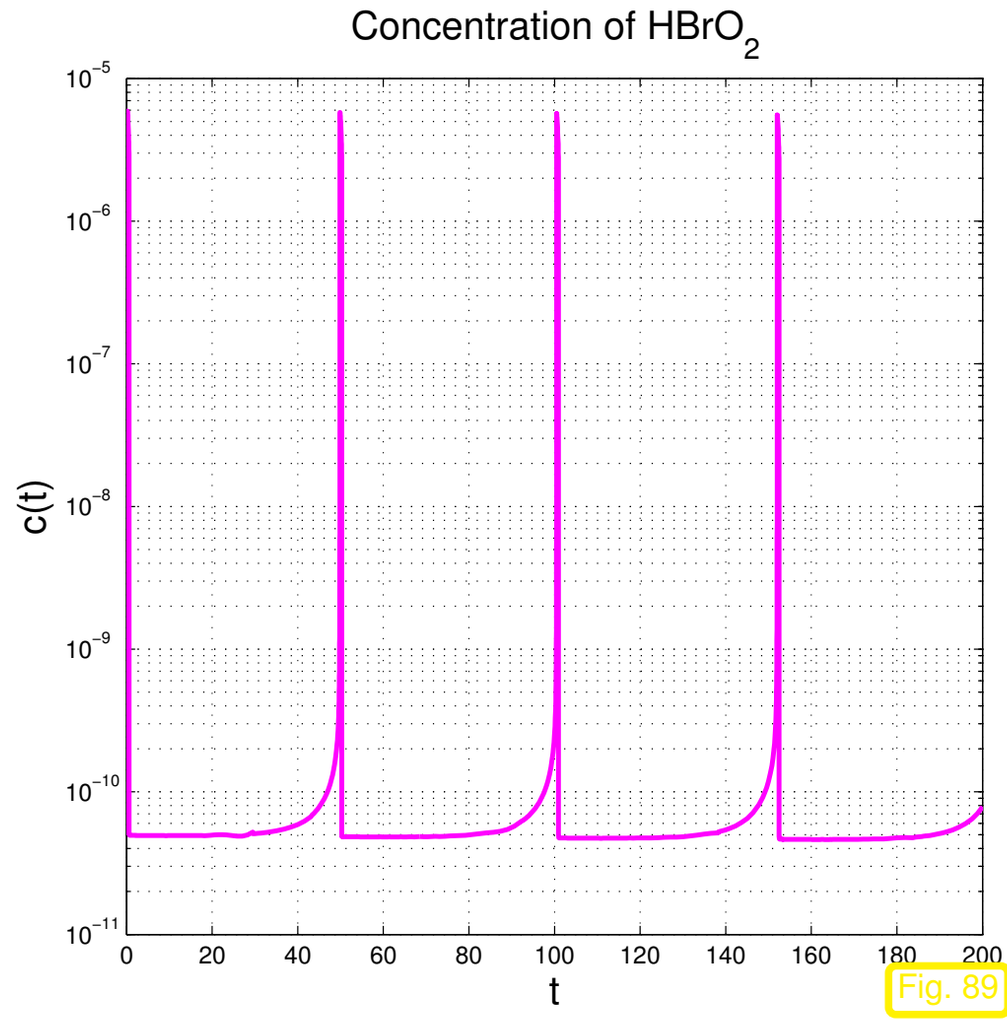
We note that the `ode45` succeeds in steadily reducing the step size as it approaches the pole of the solution.

◇

Remark 2.6.2 (Temporarily nonuniform behavior of solutions).



Kepler problem of Ex. 2.4.19



Oregonator reaction of Ex. 1.2.12

Often: solutions of IVPs show strongly nonuniform behavior w.r.t. time.

We were already introduced to a possibility of adapting a single step method to the time-local behavior of the solution:

$\rightarrow$ order control for extrapolation methods, Sect. 2.4.5

However, this does not help at all in the case of a blow-up!

$\triangle$

Fundamental question:

How to choose a *suitable* time grid $\mathcal{G} = \{t_0 < t_1 < \cdots < t_N = T\}$
for a given single step method and initial value problem?

What does *suitable* mean?

efficiency accuracy

Goal: N as small as possible & $\max_{k=1,\dots,N} \|\mathbf{y}(t_k) - \mathbf{y}_k\| < \mathtt{TOL}$ or $\|\mathbf{y}(T) - \mathbf{y}_N\| < \mathtt{TOL}$, $\mathtt{TOL}$ = tolerance

Strategy: Control the one step error by

- adjusting the *current* step size h_k,
- predicting the *next* step size h_{k+1}

} Time-local step size control

Approach: Time-local estimate of the one step error (consistency error, Def. 2.1.11) (a posteriori, as $\mathbf{y}_k, h_{k-1}$ are used)

P. Grohs

rev 63606, February 17, 2014

Why do we use *time-local* step size control, which is only based on estimating the one step error**?**

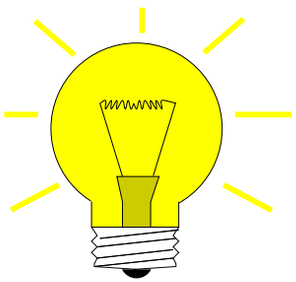
Objection: If a small error at a certain time leads to a larger error $\|\mathbf{y}_k - \mathbf{y}(t_k)\|$ at a later time, then the time-local step size control can not resolve this problem! ➢ Ex. 2.6.9

Nevertheless, the time-local step size control seems to be the only practical method,

☞ because one does not want to take many time steps backwards in time during the computation (as this results in additional computation time)

☞ because it is easy to implement and there is little computational overhead,

☞ because, as a matter of principle, one will not find a method which produces results with a guaranteed accuracy.



Idea: *Estimation of the consistency error*

Comparison of two discrete evolutions $\mathbf{\Psi}^{t,t+h}$, $\widetilde{\mathbf{\Psi}}^{t,t+h}$ of different order ($\rightarrow$ Def. 2.1.13) for the *current step size* h:
If order$(\widetilde{\mathbf{\Psi}}) >$ order$(\mathbf{\Psi})$

$$\Rightarrow \quad \underbrace{\mathbf{\Phi}^{t,t+h}\mathbf{y}(t_k) - \mathbf{\Psi}^{t,t+h}\mathbf{y}(t_k)}_{\text{consistency error}} \approx \mathtt{EST}_k := \widetilde{\mathbf{\Psi}}^{t,t+h}\mathbf{y}(t_k) - \mathbf{\Psi}^{t,t+h}\mathbf{y}(t_k) \,. \tag{2.6.3}$$

Heuristic for specific h

Example 2.6.4 (Quality of error estimate).

- Scalar IVP: $\dot{y} = \cos^2(ay)$, solution $y(t) = {}^1\!/_a \arctan(at)$ on $[-1,1]$, $a = 10$

- $\Psi \leftrightarrow$ explicit Euler method (1.4.2), order $p = 1$
  $\widetilde{\Psi} \leftrightarrow$ explicit trapezoidal rule (2.3.3), order $p = 2$

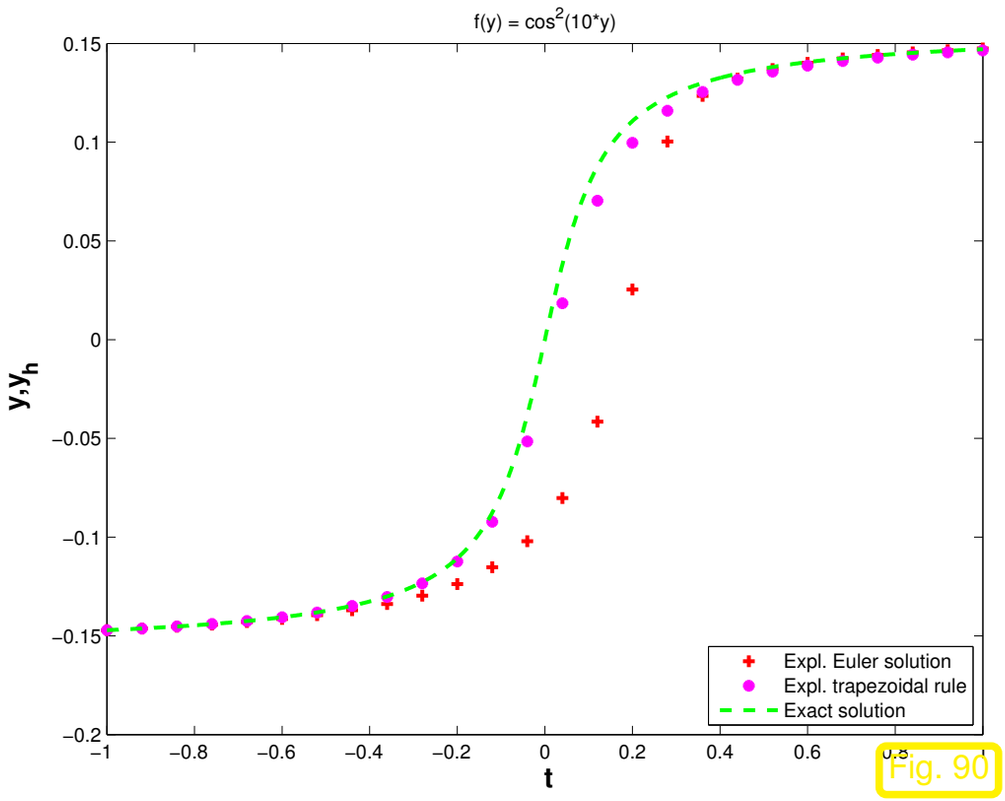


Fig. 90

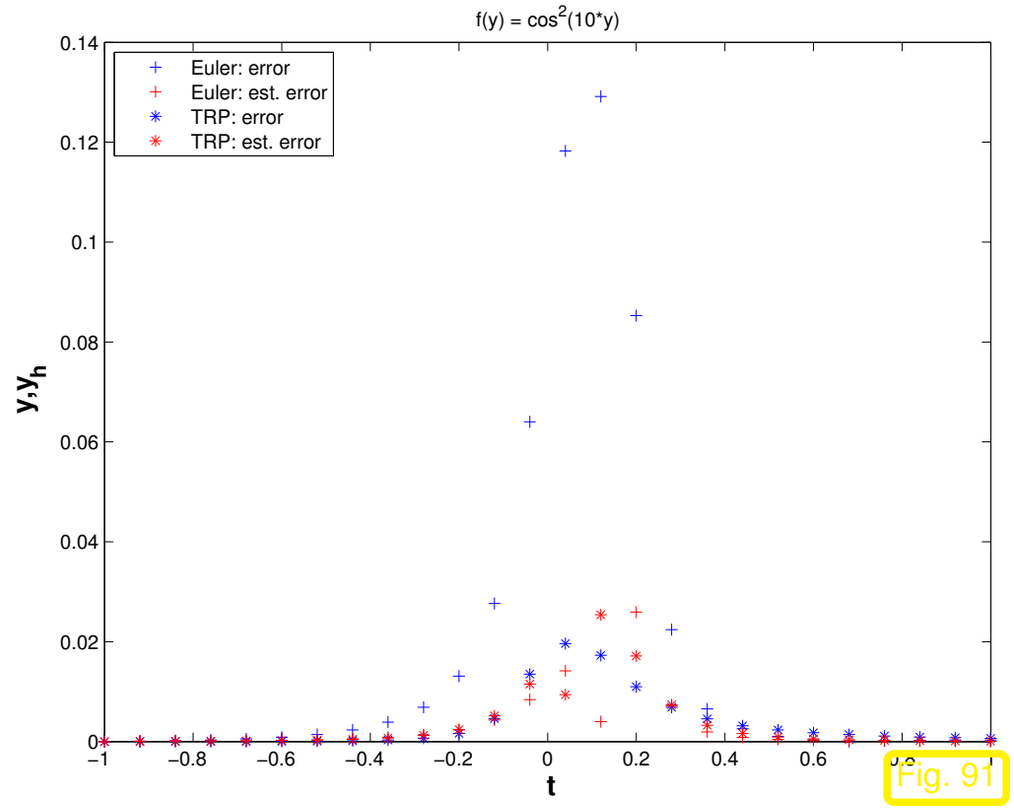


Fig. 91

Observation: ☞ Big differences between estimated and true error possible
☞ However: error estimation for $\widetilde{\Psi}$ by Ψ make sense, as "correct tendency"

Comparison $\mathtt{EST}_k \leftrightarrow \mathtt{TOL}$ ← absolute tolerance

$\mathtt{EST}_k \leftrightarrow \mathtt{TOL}\,\|\mathbf{y}_k\|$ ← relative tolerance

➢ dismiss/accept current step

Leads to a very simple algorithm:

$\mathtt{EST}_k < \mathtt{TOL}$: Execution of the current step (step size h)
Next step with step size αh, with a $\alpha > 1$ $(*)$

$\mathtt{EST}_k > \mathtt{TOL}$: Repetition of the current step with step size $< h$, e.g., $\frac{1}{2}h$

Explanation for $(*)$: If the current step size already guarantees a sufficiently small one step error, then it might be possible to also obtain a sufficiently small one step error with a somewhat bigger step size. This way, one can reduce the number of total steps, which increases the efficiency of the method. The risk of a reduction in accuracy is bounded by the error estimation in the next step.

Listing 2.4: Simple time-local step size control for single step method (autonomous ODE)

```
function [t,y] =
  odeintadapt(Psilow,Psihigh,T,y0,h0,reltol,abstol,hmin)
t = 0; y = y0; h = h0;                    %
while ((t(end) < T)  (h > hmin)) %
  yh = Psihigh(h,y0);   % ESV hoher Ordnung
```

```
  yH = Psilow(h,y0);    % ESV niedriger Ordnung
  est = norm(yH-yh);    %

  if (est < max(reltol*norm(y0),abstol))     %
    y0 = yh; y = [y,y0]; h = 1.1*h;          % Schritt akzeptiert
    h = min(h,T-t(end)); t=[t,t+h];          %
  else, h = h/2; end                         % Schritt verworfen
end
```


Comments concerning code 2.4:

- Arguments of `odeintadapt`:
  - `Psilow`, `Psihigh`: Function handles for discrete evolutions (for autonomous ODE) of different orders, type `@(y,h)`, state vector as first argument, step size as second,
  - `T`: final time $T > 0$,
  - `y0`: initial state $\mathbf{y}_0$,
  - `h0`: step size h_0 for the first time step

- `reltol`, `abstol`: Relative and absolute tolerances, compare above,
- `hmin`: minimum step size, method stops if $h_k < h_{\min}$ – this is important for identifying blow-ups and collapses.

- line 3: Check whether the final time has been reached or whether the method got stuck ($h_k < h_{\min}$).
- line 4, 5: Evolve the current state using both single step methods.
- line 6: Compute the norm of the estimated error, compare (2.6.3).
- line 8: Comparison in order to decide whether the current step should be accepted or dismissed.
- line 9, 10: Step accepted: Update the state and propose 1.1 times the current step size for the next step.
- line 11 Step dismissed: Try it again using half the step size.
- Return values
  - `t`: time grid $t_0 < t_1 < t_2 < \ldots < t_N < T$, where $t_N < T$ indicates a premature stop (collapse, blow-up),
  - `y`: sequence of states $(\mathbf{y}_k)_{k=0}^N$.

! According to our heuristic, compare (2.6.3), it seems that EST_k measures the one step error of the single step method of lower order, and that therefore we should set $\mathbf{y}_{k+1} = \mathbf{\Psi}^{h_k}\boldsymbol{y}_k$ in case the step is accepted.

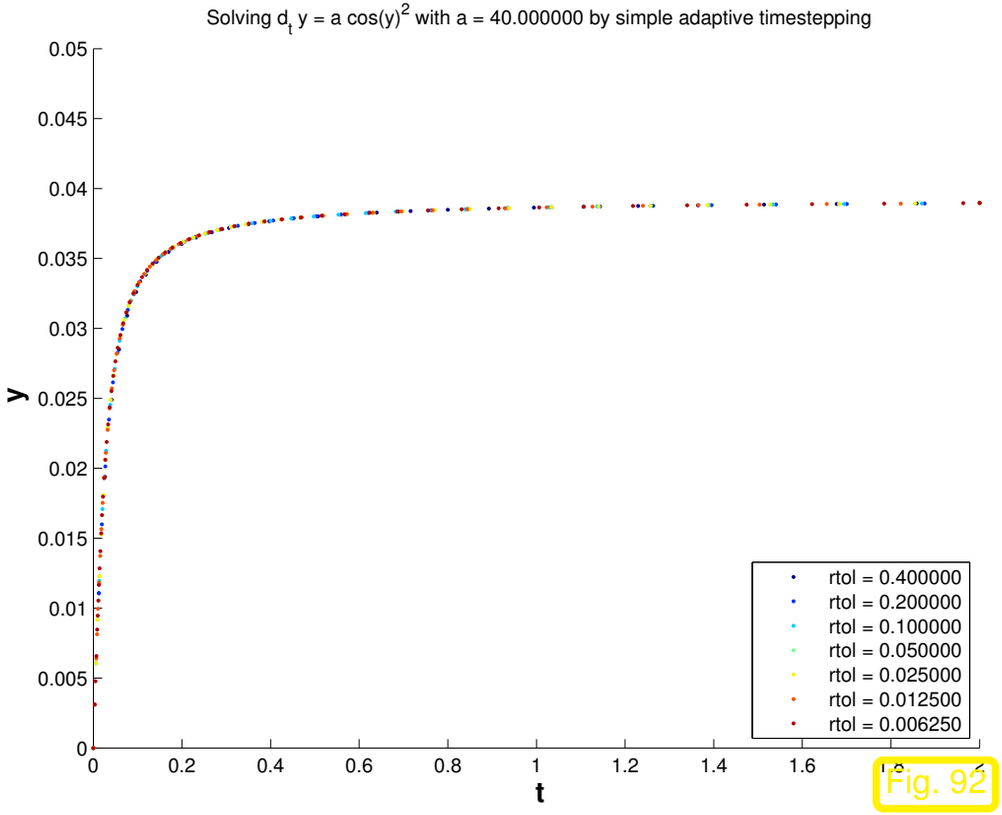
However, it makes more sense to use the most likely better approximation $\mathbf{y}_{k+1} = \widetilde{\mathbf{\Psi}}^{h_k}\boldsymbol{y}_k$, which is available without any additional effort. Every implementation of time-local step size control uses this approach, in particular the code 2.4. This approach can be justified by arguments from control theory, compare as well the following Rem. 2.6.7.

Example 2.6.5 (Gain in efficiency by adaptivity). $\rightarrow$ Ex. 2.6.4

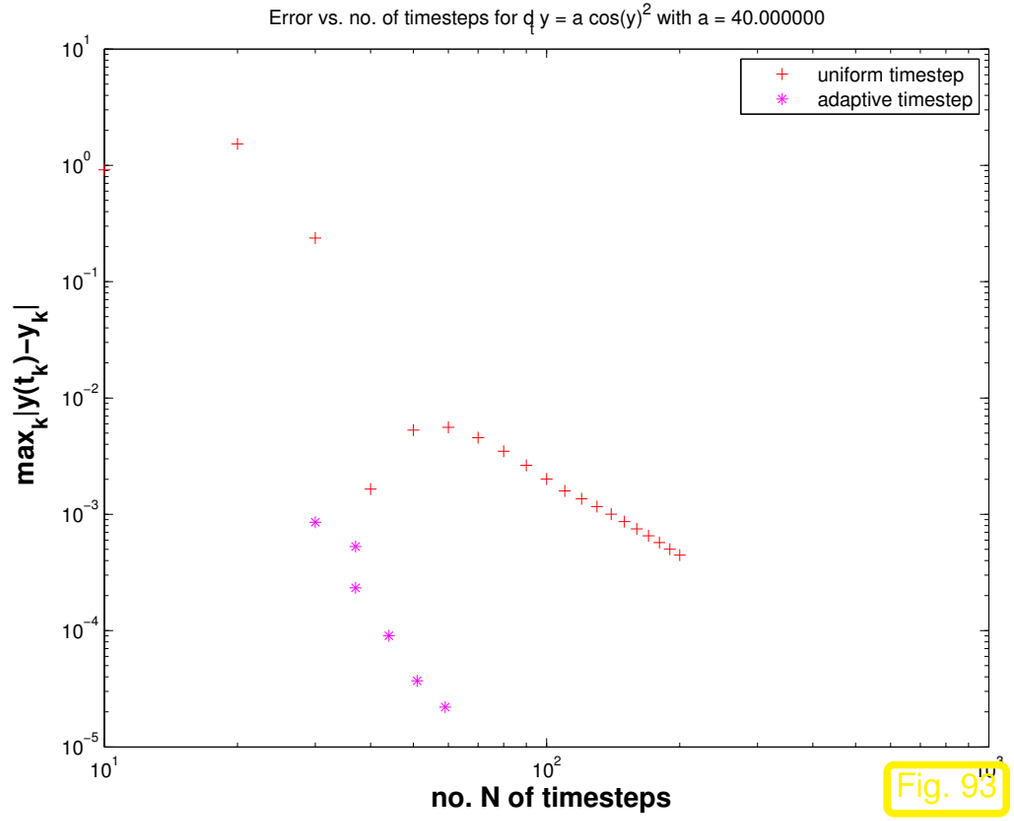
- IVP for ODE $\dot{y} = \cos(\alpha y)^2$, $\alpha > 0$,
- Analytic solution $y(t) = \arctan(\alpha(t-c))/\alpha$ for $y(0) \in]-\pi/2, \pi/2[$
- Integration interval $[0,2]$, initial value $y(0) = 0$

Simple adaptive strategy from code 2.4 with the local error estimate from Ex. 2.6.4.

Now we examine in how far the discretization error depends on the computational work, which is proportional to the number of time steps.



Solution $(\mathbf{y}_k)_k$ for different `rtol`



Error as function of computational work

P. Grohs

rev 63606, February 17, 2014

☞ For comparable computational work, the adaptive step size control produces significantly more accurate results than the same single step method with uniform step size.

Disadvantage of code 2.4: Cross-the-board amplification/reduction of the step sizes in lines 9, 11 "wastes" information contained in $\mathrm{EST}_k : \mathtt{TOL}$.

We want more **!**

If $\mathtt{EST}_k > \mathtt{TOL}$: step size correction $t_{k+1} =$ **?**
If $\mathtt{EST}_k < \mathtt{TOL}$: step size proposition $t_{k+2} =$ **?**

step size correction refers to a dismissed, to be repeated *current* step
step size proposition is used for the *next* step
(cmp. code 2.4)

If order$(\boldsymbol{\Psi}) = p$, order$(\widetilde{\boldsymbol{\Psi}}) > p$, $p \in \mathbb{N}$,

goal: efficiency

$$\begin{aligned} \boldsymbol{\Psi}^{t,t+h}\mathbf{y}(t_k) - \boldsymbol{\Phi}^{t,t+h}\mathbf{y}(t_k) &= ch^{p+1} + O(h^{p+2}) \,, \\ \widetilde{\boldsymbol{\Psi}}^{t,t+h}\mathbf{y}(t_k) - \boldsymbol{\Phi}^{t,t+h}\mathbf{y}(t_k) &= O(h^{p+2}) \end{aligned} \quad \overset{h \ll 1}{\Rightarrow} \quad \mathtt{EST}_k \approx ch^{p+1} \overset{!}{=} \mathtt{TOL} \,.$$

heuristic!

► "Optimal step size": (step size proposition)

$$\boxed{h^* = h\sqrt[p+1]{\frac{\mathtt{TOL}}{\mathtt{EST}_k}}} \quad . \qquad (2.6.6)$$

Corrected step size

step size proposition

Remark 2.6.7 (Control of $\widetilde{\Psi}$ by Ψ). Code 2.4

Consideration so far: "Error estimate" of $\widetilde{\Psi}$ ➢ control of $\widetilde{\Psi}$

Efficient **?** More accurate (more expensive) method $\widetilde{\Psi}$ is only used to control the less accurate method.

With the same work: Integration of IVP with $\widetilde{\Psi}$ controlled by Ψ !

This is how it is done in practice !

Reminder of Ex. 2.6.4: Euler method (order $p = 1$) produced good error estimate for explicit trapezoidal rule (order $p = 2$)

Another heuristic:
$\mathrm{EST}_k > \mathrm{TOL}$ is a hint that one of the two methods $\boldsymbol{\Psi}$, $\widetilde{\boldsymbol{\Psi}}$ has problems with the (local) approximation of the solution. Therefore, a reduction of h_k is indicated.

Mathematical justification: control theory $\rightarrow$ [12, Sect. 5.2]

$\triangle$

MATLAB implementation:

- $\Psi, \widetilde{\Psi} \hat{=}$ discrete evolutions, order of consistency $p/p+1$
- $t_0 \hat{=}$ initial time, $T \hat{=}$ final time
- $\mathbf{y}_0 \hat{=}$ initial value (column vector)
- `reltol`, `abstol` $\hat{=}$ absolute/relative tolerances
- $h_0, h_{\min} \hat{=}$ step size for 1st step/minimum step size

SSM with step size control

```
function [t,y] = ssctrl(Ψ,Ψ̃,t0,T,y0,h0,reltol,abstol,hmin)
t = t0; y = y0; h = h0;
while ((t(end) < T) && (h > hmin))
  yh = Ψ̃(t(end),y(:,end),h);
  yH = Ψ(t(end),y(:,end),h);
  est = norm(yH-yh);
  tol = max(reltol*norm(y(:,end)),abstol);
  h = h*max(0.5,min(2,(tol/est)^(1/(p+1))));
  if (est < tol)
    y = [y,yh]; h = min(h,T-t(end)); t = [t,t(end)+h];
  end
end
```

P. Grohs

rev 63606, February 17, 2014

Example 2.6.8 (Step size control for explicit trapezoidal rule/Euler method).

Initial value problem for scalar logistic differential equations, compare Ex. 1.2.1

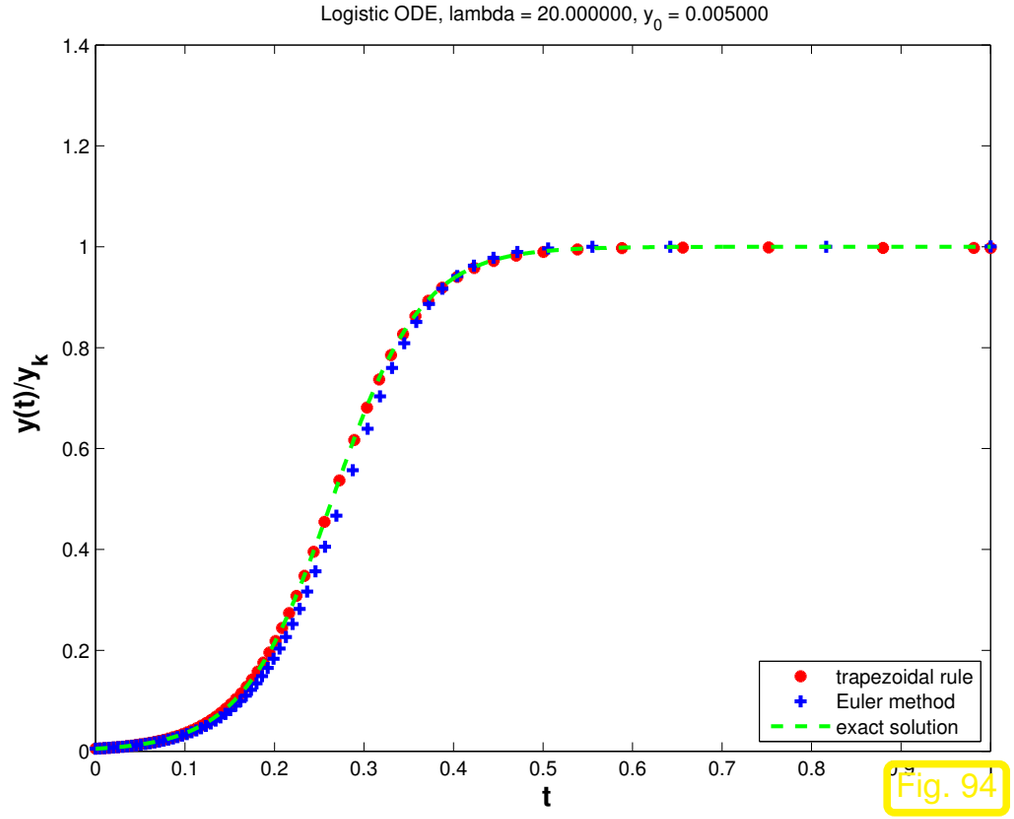
$$\dot{y} = \lambda y(1-y)\,, \quad \lambda = 20 \quad \rightarrow \quad y(t) = \frac{y_0}{y_0 + (1-y_0)\exp(-\lambda t)}\,.$$

Single step method from Ex. 2.6.4, adaptation of step size according to (2.6.6)

❶ Integration with explicit Euler method (1.4.2), error estimate (2.6.3) with explicit trapezoidal rule (2.3.3)

❷ Integration with explicit trapezoidal rule (2.3.3), step size control with explicit Euler method according to Rem. 2.6.7

Absolute/relative tolerances = 0.005, $y_0 = 0.1/\lambda$



Trapezoidal rule/Euler: 63/62 steps, 12 dismissed

P. Grohs

rev 63606, February 17, 2014

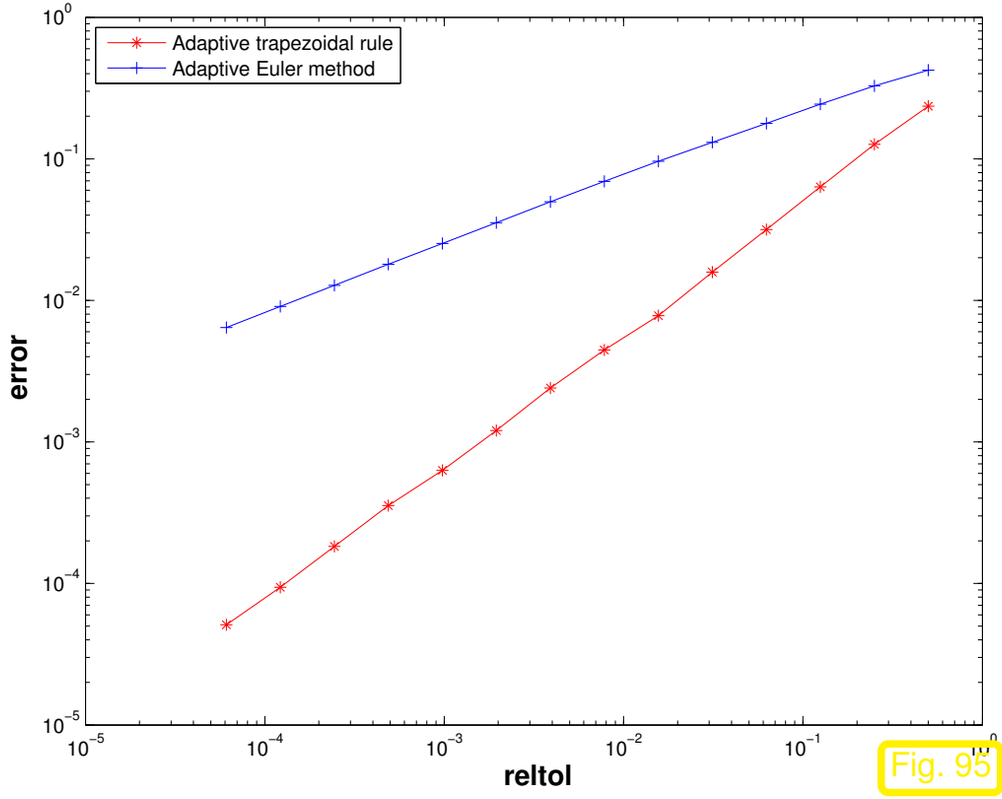


Fig. 95

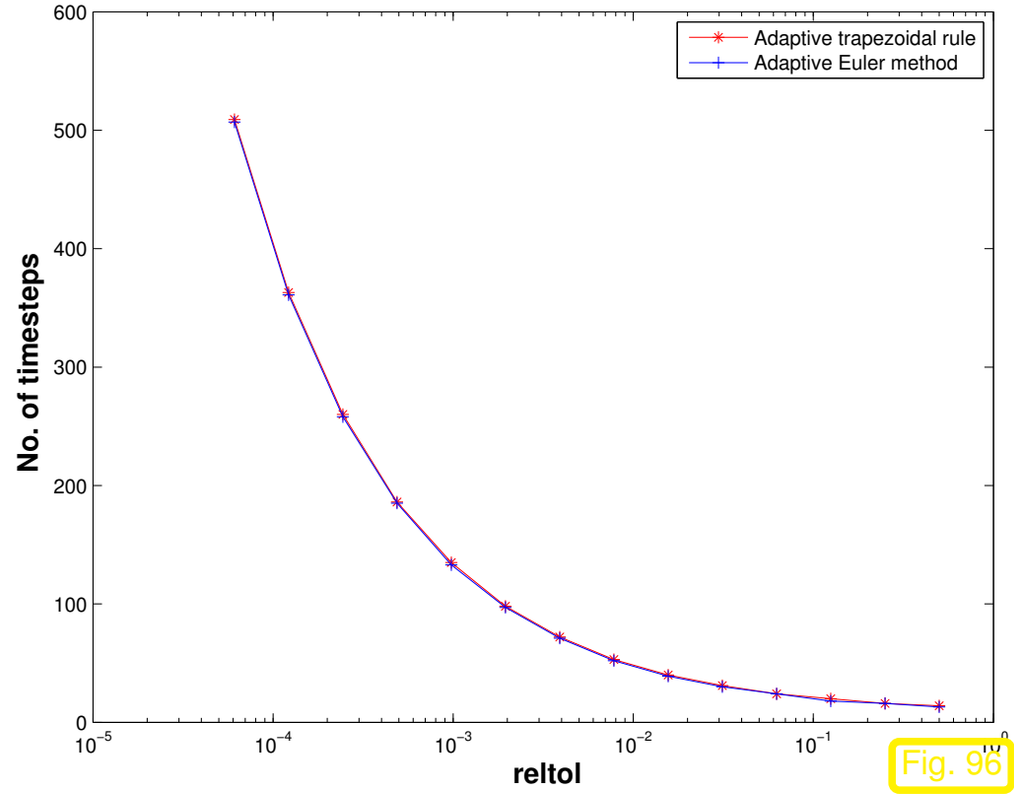


Fig. 96

P. Grohs

rev 63606, February 17, 2014

In this example, the approach of using the approximation generated by the higher order method $\widetilde{\Psi}$ to continue the computation clearly results in higher accuracy.

Observation: for this example, the error $\max_j |y(t_j) - y_j|$ is well correlated with the tolerance TOL.

$\diamond$

Example 2.6.9 (“Failure” of adaptive time step control). $\rightarrow$ Ex. 2.6.5

Same ODE and simple adaptive time step control as in Ex. 2.6.5. Same evaluations as well.

Now: initial value $y(0) = -0.0386$, cmp. Ex. 2.6.4.

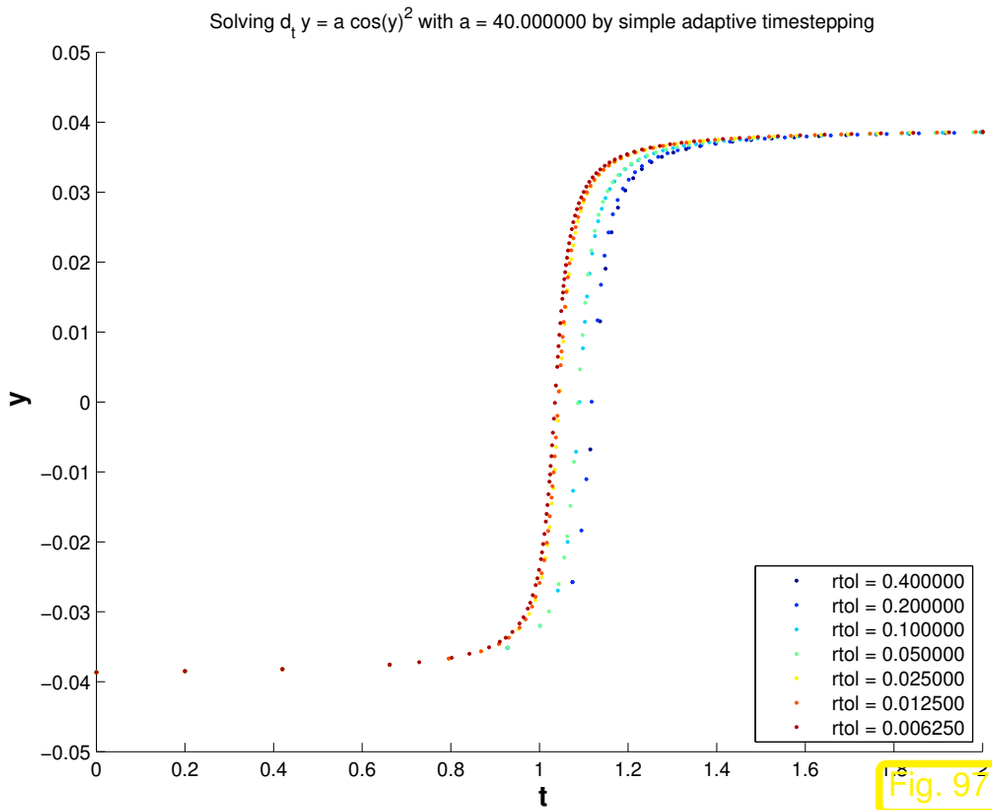


Fig. 97

Solutions $(\mathbf{y}_k)_k$ for different `rtol`

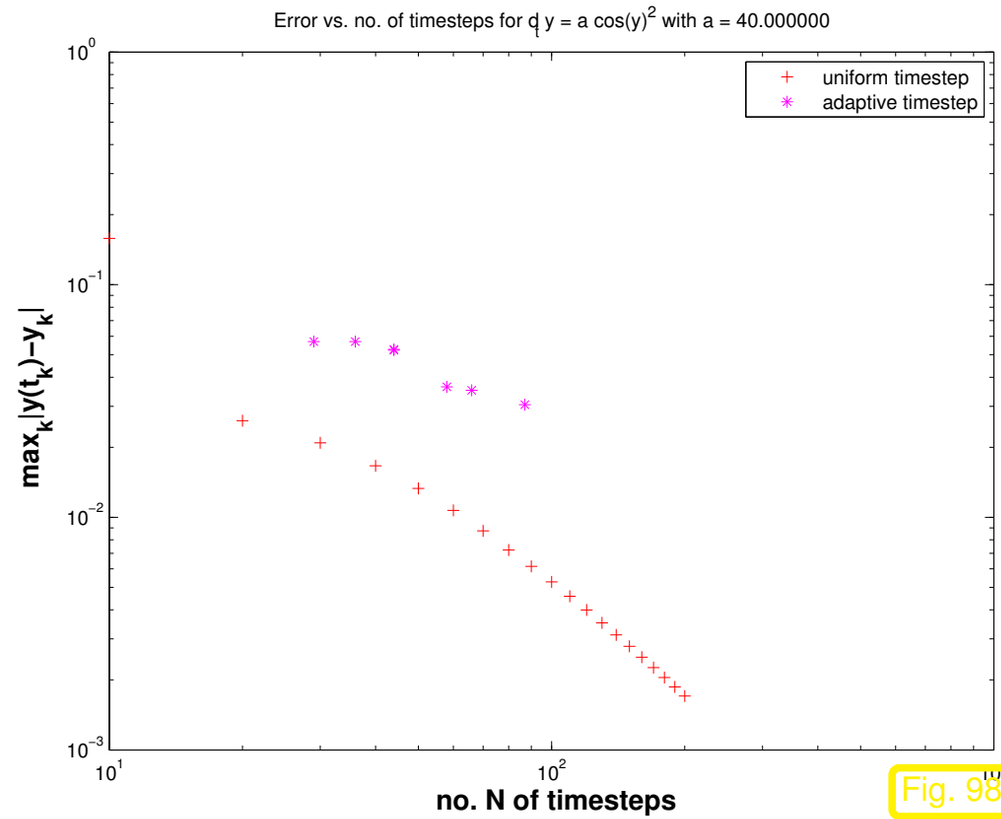


Fig. 98

Error as function of computational work

☞ Bigger error for adaptive time step control than for using uniform step size

Explanation: the position of the steep slopes depends *sensitively* on the initial data. Therefore, small one step errors in the first few time steps will result in big errors at time $t \approx 1$. The local step size control considers these small initial one step errors to be negligible. Therefore, it cannot prevent the significant discretization errors at later time generated by the small errors at the initial time.

General context: in the case of *ill-conditioned* initial value problems (i.e., the solution depends sensitively on the initial values, cmp. Sect. 1.3.3.5, “chaotic systems”), even a tiny one step error in the first step can result in a discrete solution that is completely different from the exact solution. However, for such problems, the concept of defining the accuracy based on the discretization error is not appropriate, compare the discussion in Sect. 1.3.3.5.

◇

Example 2.6.10 (Step size control and instability).

- Initial value problem for scalar logistic diff. eqn., compare Ex. 2.6.8, now $\lambda = 100$
- Explicit Euler method (1.4.2), explicit trapezoidal rule (2.3.3) with step size control as in Ex. 2.6.8

Absolute/relative tolerance = 0.05, initial step size (for adaptive SSM) $h = 0.05$

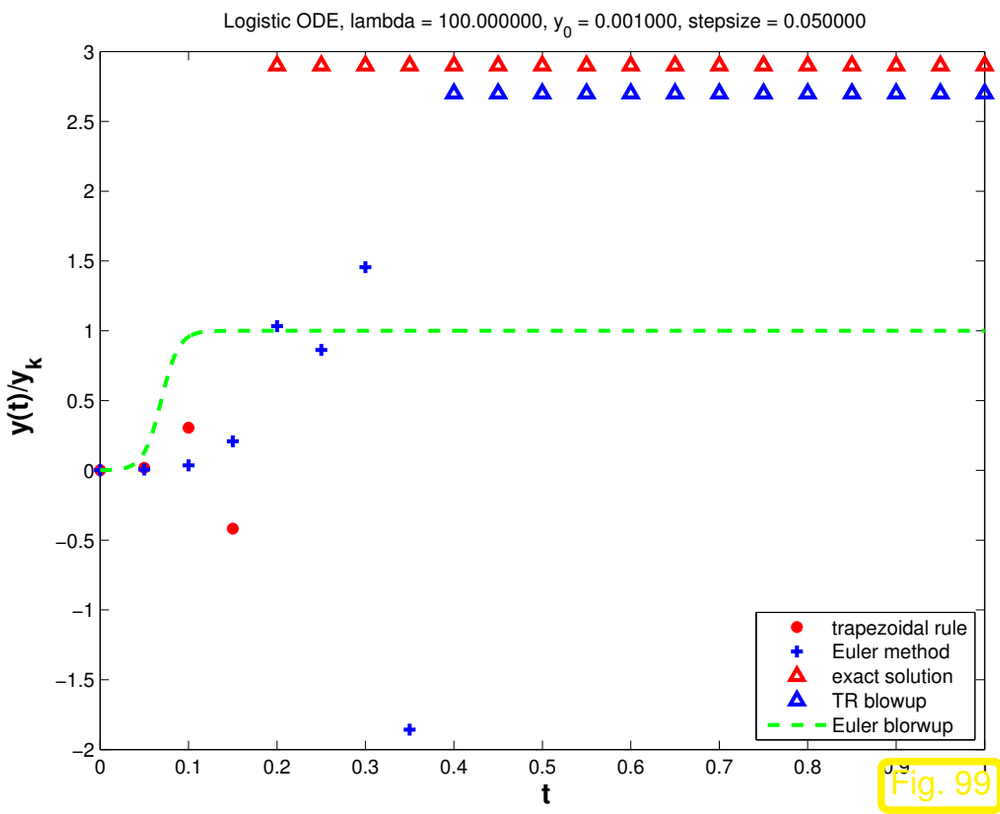


Fig. 99

Trapezoidal/Euler: uniform step size $h = 0.05$

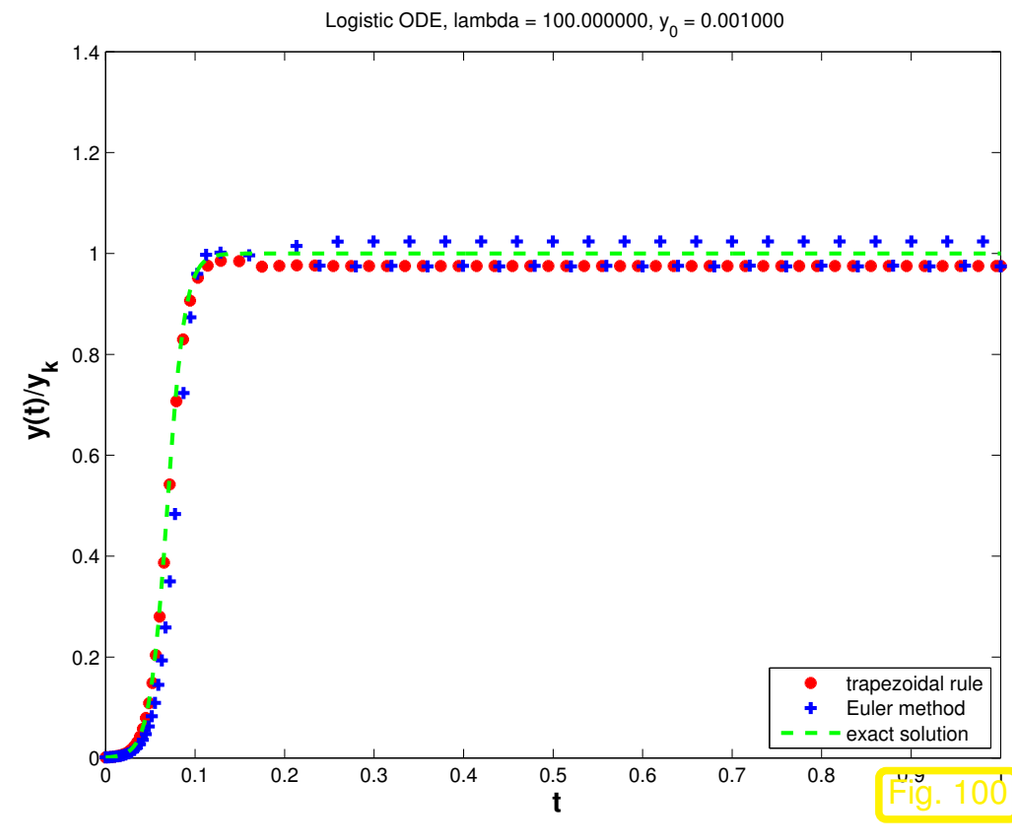


Fig. 100

Trapezoidal/Euler: 119/114 steps, 28/42 dismissed

step size control prevents instability, cmp. Ex. 1.4.9 !

Example 2.6.11 (Step size control and collapse).

Scalar initial value problem with collapse, cmp. Ex. 1.3.11

$$\dot{y} = -\frac{1}{\sqrt{y}}\,, \quad y(0) = 1$$

$$\Rightarrow \quad y(t) = (1 - 3t/2)^{2/3}\,.$$

step size control as in Ex. 2.6.8 , absolute/relative tolerance = 0.005

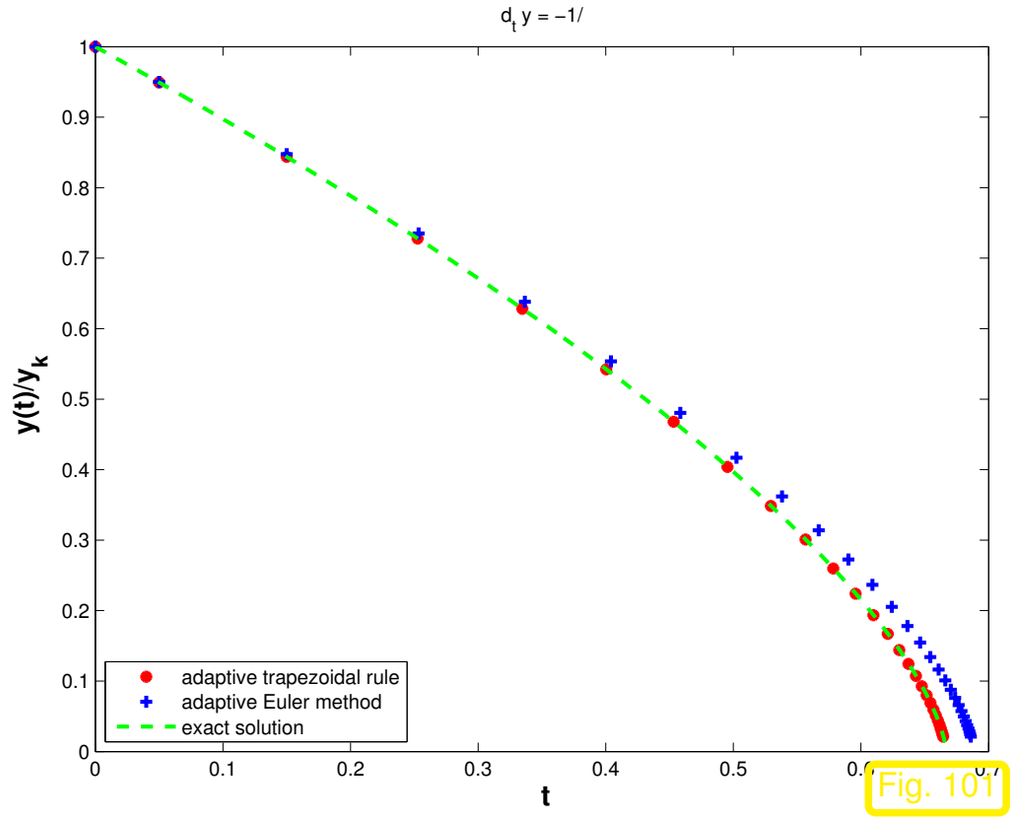


Fig. 101

step size control ➢ method “catches” collapse of the solution

Example 2.6.12 (Step size control and blow-up).

Scalar initial value problem with blow-up, cmp. Ex. 1.3.11

$$\dot{y} = y^2 \;, \quad y(0) = 1$$
$$\Rightarrow \quad y(t) = \frac{1}{1-t} \;.$$

step size control as in Ex. 2.6.8, absolute/relative tolerance = 0.05

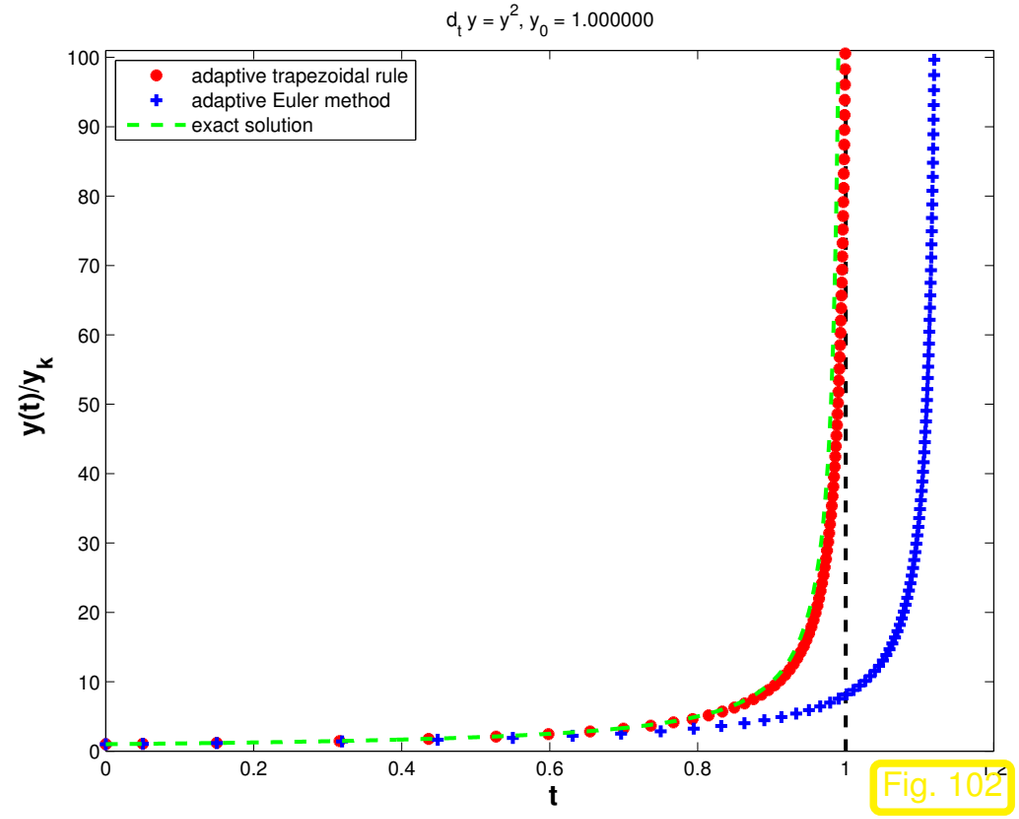


Fig. 102

step size control ➢ method “catches” blow-up of the solution

◇

Remark 2.6.13 (Embedded RK SSM).

Algorithmic realization (SSM):

Embedded Runge-Kutta method

$$\begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^T \\ \hline & \widehat{\mathbf{b}^T} \end{array} := \begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \\ \hline & \widehat{b}_1 & \cdots & \widehat{b}_s \end{array} .$$

Embedded RK SSM: Butcher tableau

Same increments $\mathbf{k}_i$, different weights b_i ($\rightarrow$ Def 2.3.5) realize RK evolutions $\mathbf{\Psi}_h, \widetilde{\mathbf{\Psi}}_h$ of orders p and $p+1$.

$$\mathbf{\Psi}_h\mathbf{y} = \mathbf{y} + h\sum_{i=1}^{s} b_i\mathbf{k}_i \quad , \quad \widetilde{\mathbf{\Psi}}_h\mathbf{y} = \mathbf{y} + h\sum_{i=1}^{s} \widehat{b}_i\mathbf{k}_i \ .$$

Motivation: efficiency (increments $\mathbf{k}_i$ only need to be computed once, compare Def. 2.3.5)

Common: $p = 4$, $p = 7$

$\triangle$

Example 2.6.14 (Embedded Runge-Kutta methods). $\rightarrow$ [22, Sect. II.4]

$$
\begin{array}{c|ccccc}
0 & & & & & \\
\frac{1}{3} & \frac{1}{3} & & & & \\
\frac{1}{3} & \frac{1}{6} & \frac{1}{6} & & & \\
\frac{1}{2} & \frac{1}{8} & 0 & \frac{3}{8} & & \\
1 & \frac{1}{2} & 0 & -\frac{3}{2} & 2 & \\
\hline
y_1 & \frac{1}{6} & 0 & 0 & \frac{2}{3} & \frac{1}{6} \\
\hline
\widehat{y}_1 & \frac{1}{10} & 0 & \frac{3}{10} & \frac{2}{5} & \frac{1}{5}
\end{array}
$$

Embedded RK method of order 4(“5”) by Merson

$$
\begin{array}{c|ccccc}
0 & & & & & \\
\frac{1}{2} & \frac{1}{2} & & & & \\
\frac{1}{2} & 0 & \frac{1}{2} & & & \\
1 & 0 & 0 & 1 & & \\
\frac{3}{4} & \frac{5}{32} & \frac{7}{32} & \frac{13}{32} & -\frac{1}{32} & \\
\hline
y_1 & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & \\
\hline
\widehat{y}_1 & -\frac{1}{2} & \frac{7}{3} & \frac{7}{3} & \frac{13}{6} & -\frac{16}{3}
\end{array}
$$

Embedded RK method of order 3(4) by Zonneveld

$$
\begin{array}{c|ccccccc}
0 & & & & & & & \\
\frac{1}{5} & \frac{1}{5} & & & & & & \\
\frac{3}{10} & \frac{3}{40} & \frac{9}{40} & & & & & \\
\frac{4}{5} & \frac{44}{45} & -\frac{56}{15} & \frac{32}{9} & & & & \\
\frac{8}{9} & \frac{19372}{6561} & -\frac{25360}{2187} & \frac{64448}{6561} & -\frac{212}{729} & & & \\
1 & \frac{9017}{3168} & -\frac{355}{33} & \frac{46732}{5247} & \frac{49}{176} & -\frac{5103}{18656} & & \\
1 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\
\hline
y_1 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\
\hline
\widehat{y}_1 & \frac{5179}{57600} & 0 & \frac{7571}{16695} & \frac{393}{640} & -\frac{92097}{339200} & \frac{187}{2100} & \frac{1}{40}
\end{array}
$$

DOPRI5: Embedded RK method of order 4(5) by Dormand & Prince (MATLAB `ode45`)

◇

P. Grohs

rev 63606, February 17, 2014

Adaptive integrators for initial value problems in MATLAB:

```
options = odeset('abstol',atol,'reltol',rtol,'stats','on');
[t,y] = ode45/ode23(@(t,x) f(t,x),tspan,y0,options);
```

(`f` = function handle, `tspan` $\hat{=}$ $[t_0, T]$, `y0` $\hat{=}$ $\mathbf{y}_0$, `t` $\hat{=}$ t_k, `y` $\hat{=}$ $\mathbf{y}_k$)

Example 2.6.15 (Adaptive RK SSM for the computation of particle paths). → Ex. 2.4.19

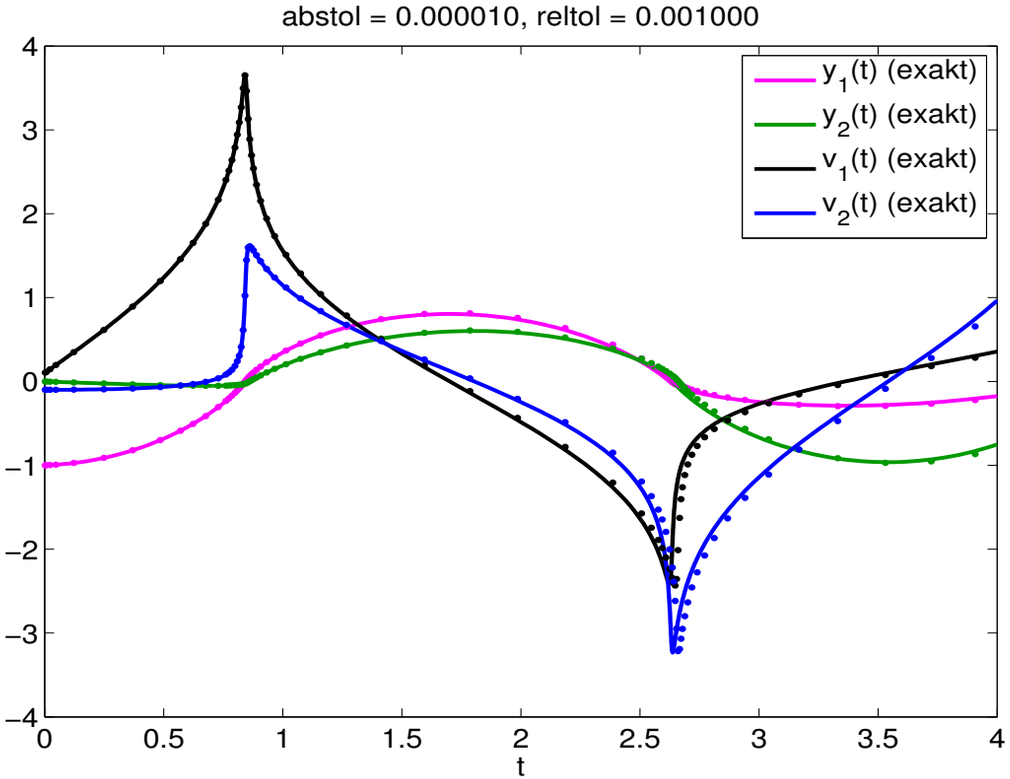
Movement of a charged particle in the field of a straight wire = line charge (conservative central field, center $\binom{0}{0}$, potential $U(\boldsymbol{x}) := -2\log\|\boldsymbol{x}\|$): → Ex. 1.2.25

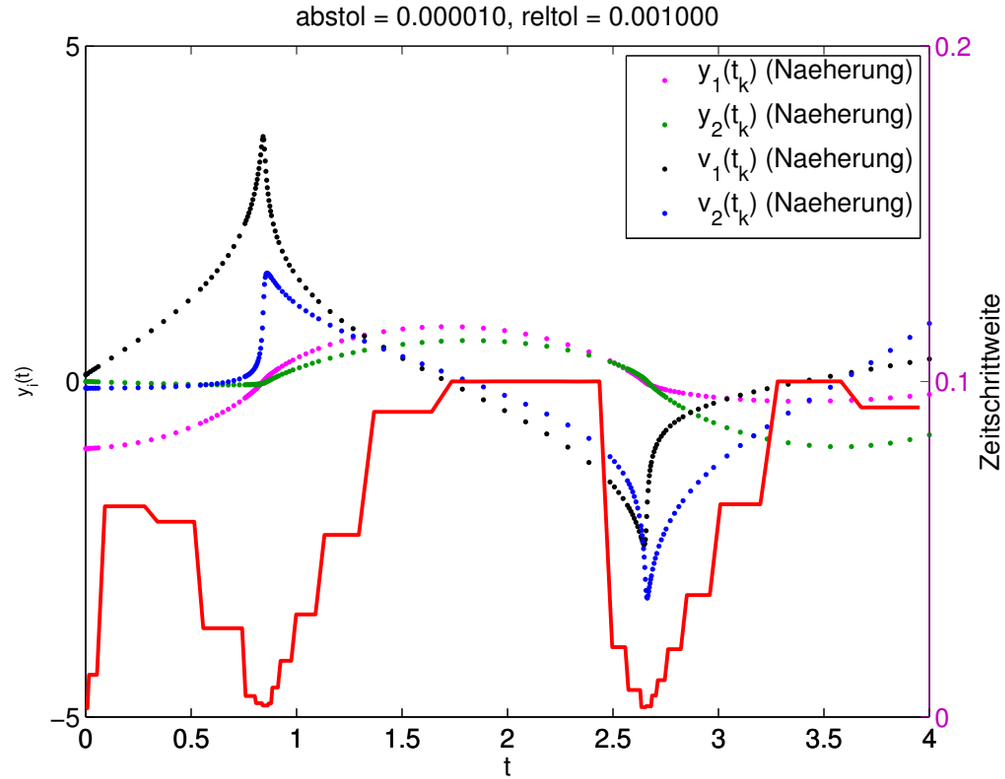
$$\ddot{\mathbf{y}} = -\frac{2\mathbf{y}}{\|\mathbf{y}\|^2} \quad \Rightarrow \quad \dot{\begin{pmatrix}\mathbf{y}\\ \mathbf{v}\end{pmatrix}} = \begin{pmatrix}\mathbf{v}\\ -\frac{2\mathbf{y}}{\|\mathbf{y}\|^2}\end{pmatrix} \quad , \quad \mathbf{y}(0) = \begin{pmatrix}-1\\ 0\end{pmatrix}, \mathbf{v}(0) = \begin{pmatrix}0.1\\ -0.1\end{pmatrix} .$$

initial value: $\mathbf{y}(0) = (-1, 0, 0.1, -0.1)$, final time: $T = 4$

Adaptive integrator: `ode45(@(t,x) satf,[0 4],[-1;0;0.1;-0.1,],options)`:

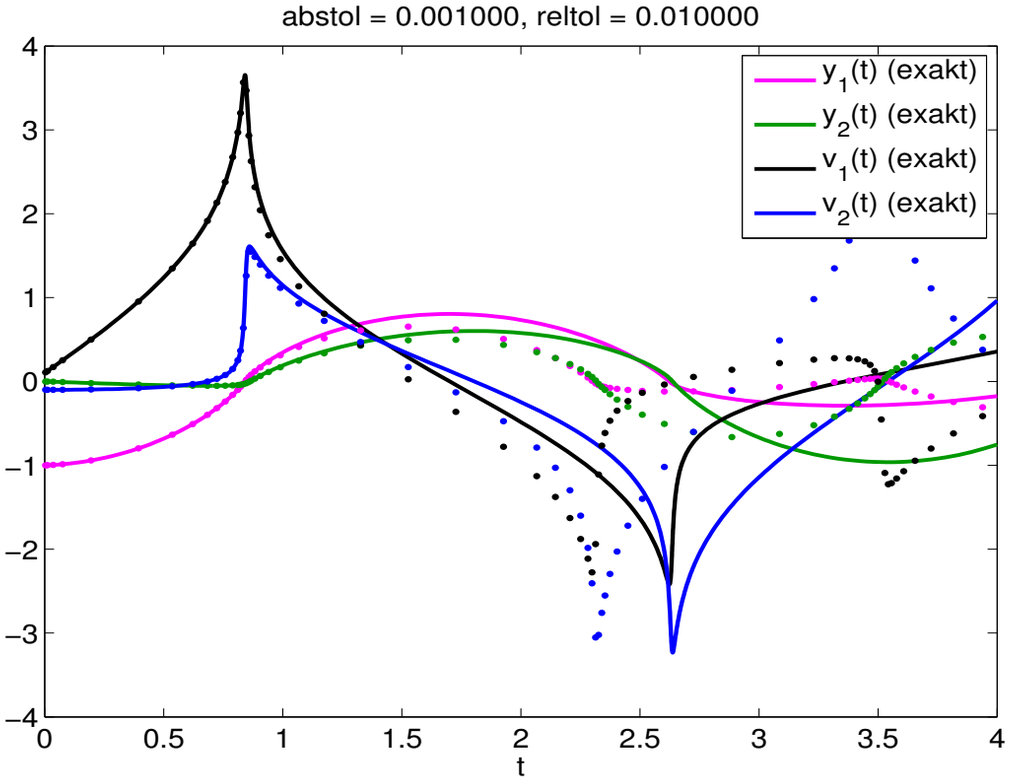
❶ `options = odeset('reltol',0.001,'abstol',1e-5);`
❷ `options = odeset('reltol',0.01,'abstol',1e-3);`

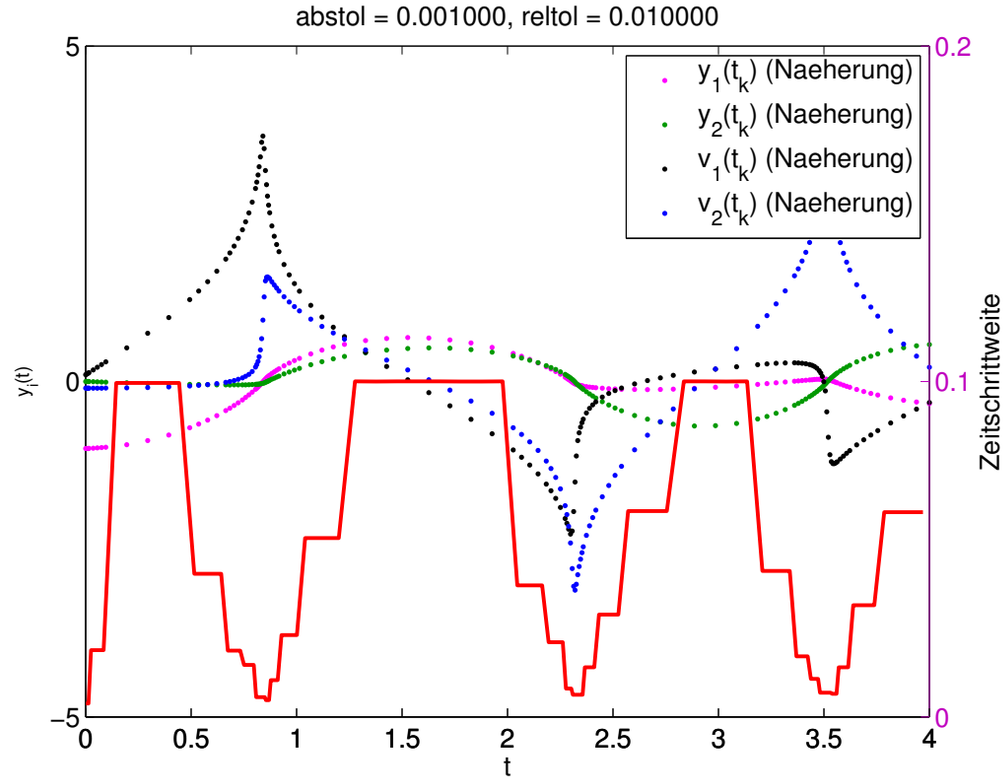




P. Grohs

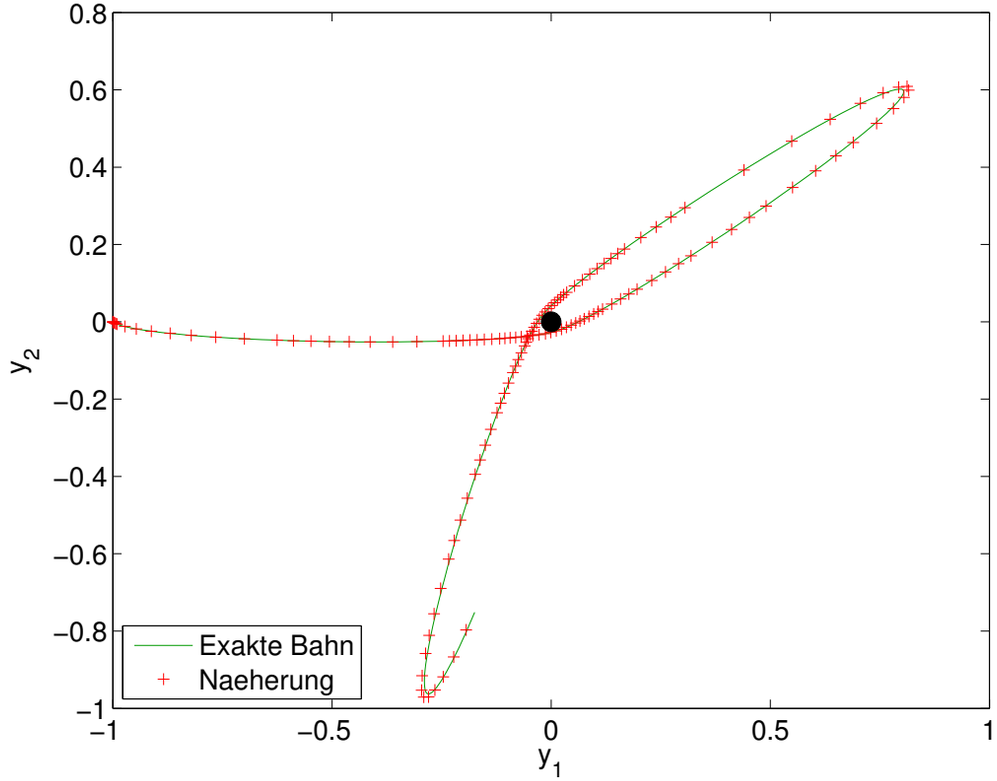
rev 63606, February 17, 2014



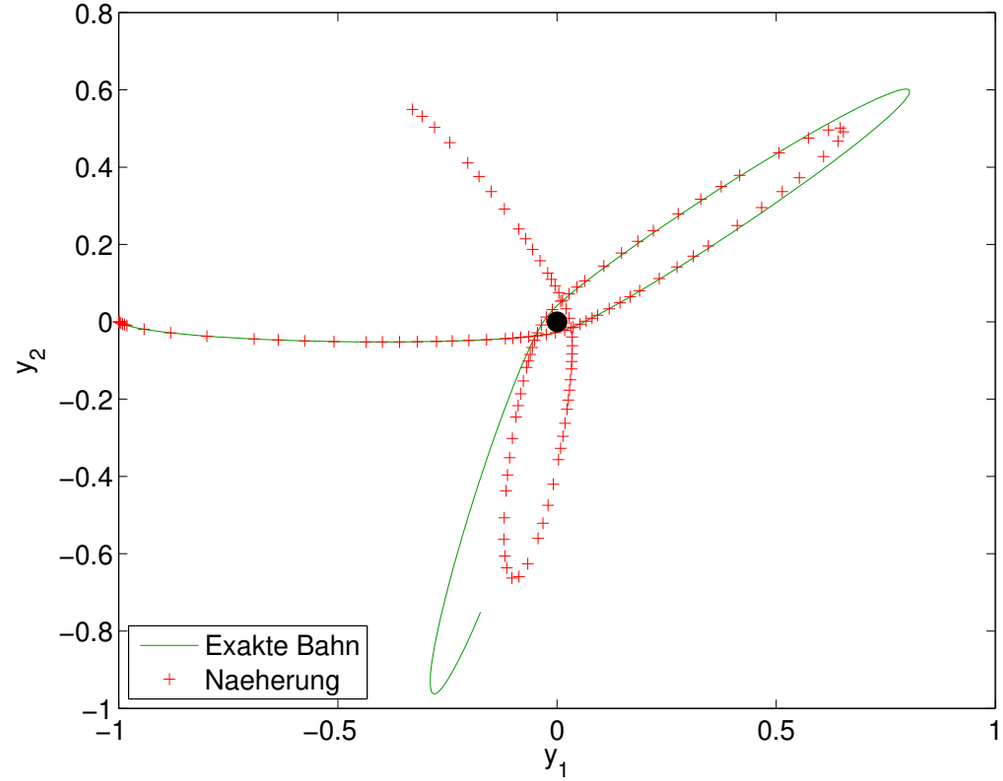


P. Grohs

rev 63606, February 17, 2014



reltol=0.001, abstol=1e-5



reltol=0.01, abstol=1e-3

P. Grohs

rev 63606, February 17, 2014

☞ Qualitatively wrong solution for slightly reduced tolerance **!**

◇

Example 2.6.16 (Step size control for equations of motion). → Bsp. 2.4.19

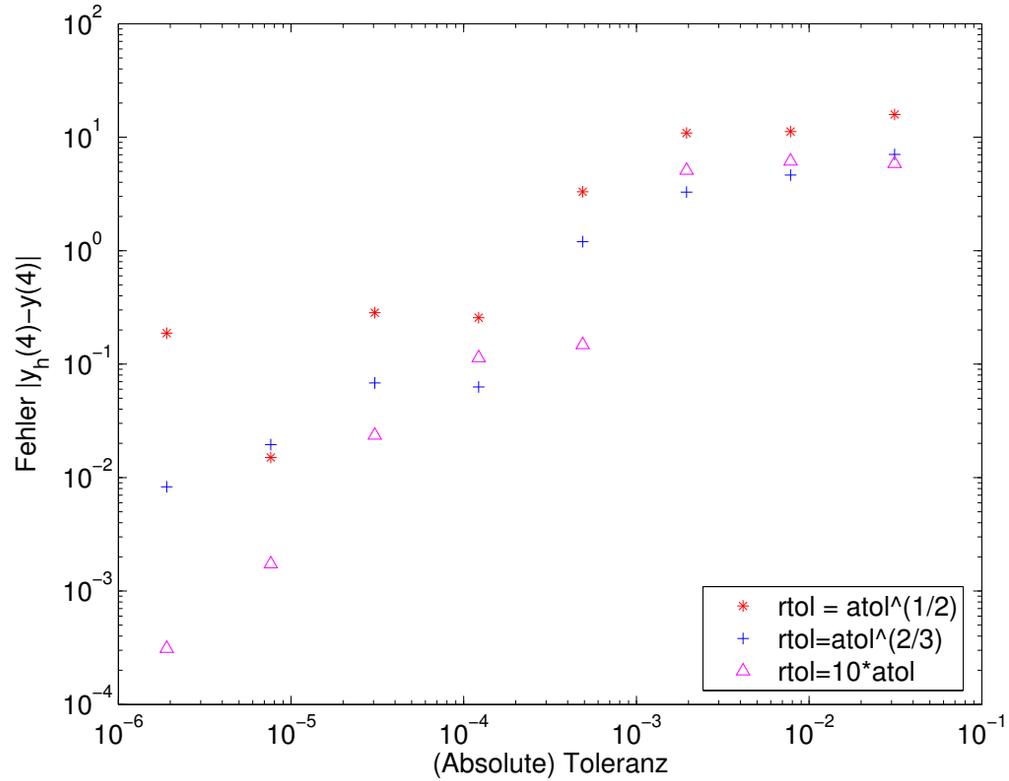
IVP from Ex. 2.4.19

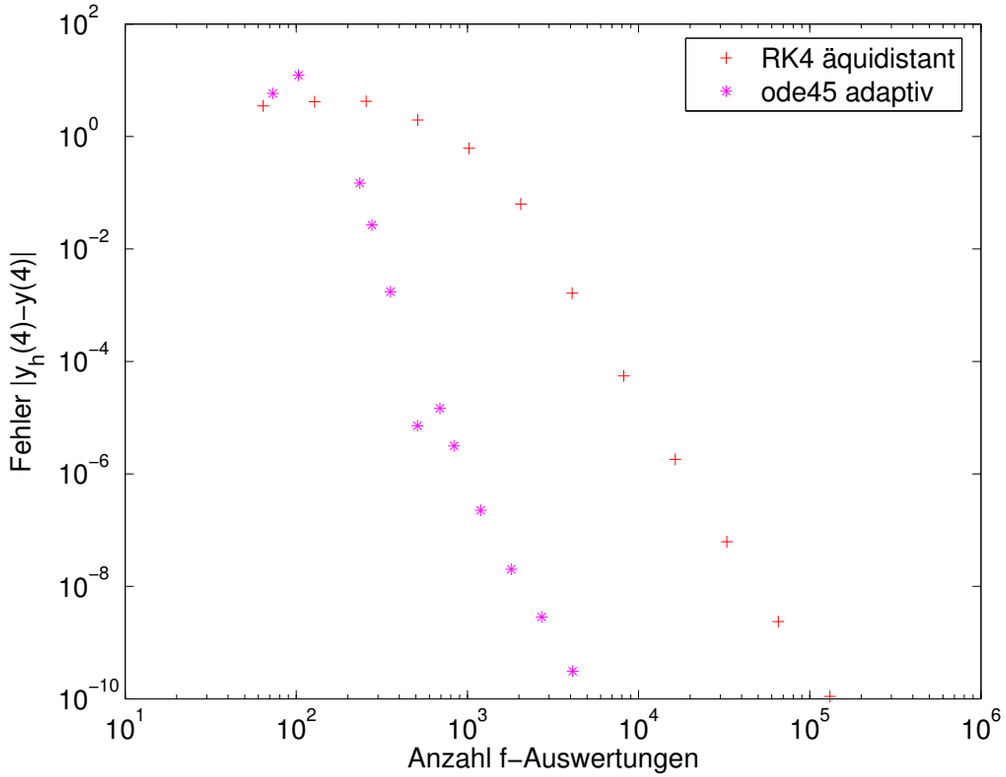
`ode45` with different absolute/relative tolerances

In contrast to Ex. 2.6.8:

! tolerances do not hold information concerning the global error

Explanation: just like in Ex. 2.6.9, this is an ill-conditioned IVP, which makes it impossible to predict the influence of one step errors on the discretization error.





Efficiency of step size control:

Comparison:

- Classic Runge-Kutta method (2.3.11)
- Embedded Runge-Kutta method with step size control: `ode45`

Measure for cost: $\sharp f$-evaluations

Adaptivity pays off !

◇

P. Grohs

rev 63606, February 17, 2014

3 Stability [12, Kap. 6]

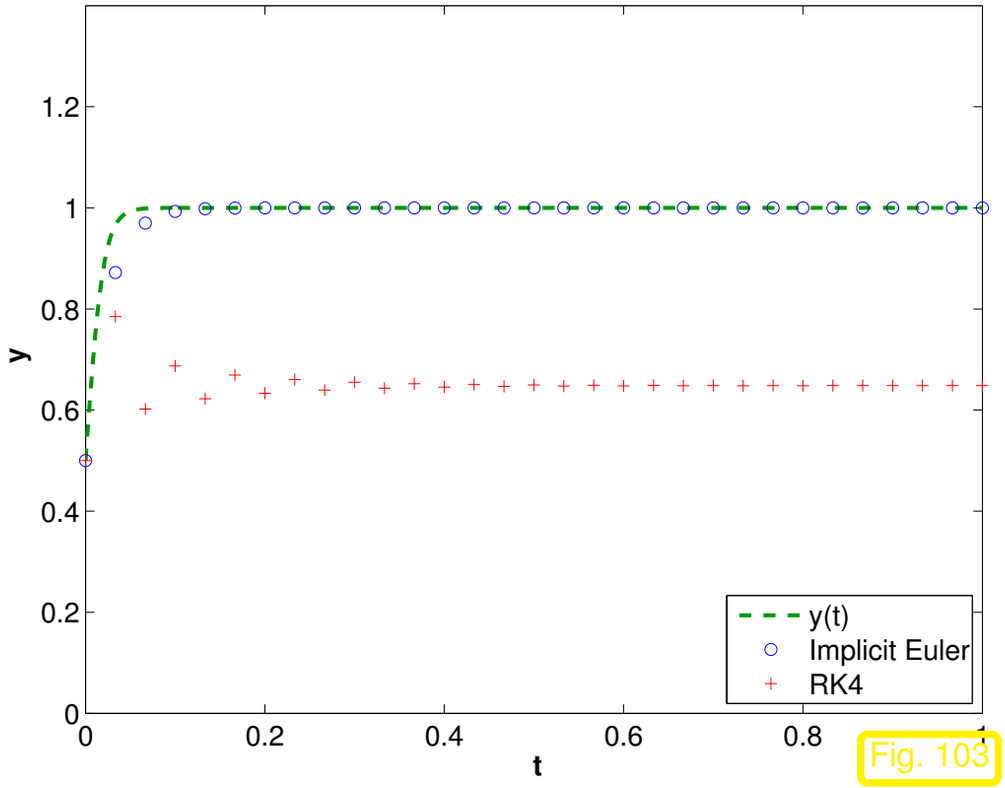
Example 3.0.1 (Inefficiency of explicit Runge-Kutta-methods). $\rightarrow$ Ex. 1.4.9, 1.4.15

logistic differential equation $\dot{y} = f(y)$, $f(y) = \lambda y(1-y) \rightarrow$ (2.2.57), $\lambda = 50$, Initial value $y_0 = 0.1$, time-interval $[0,1]$:

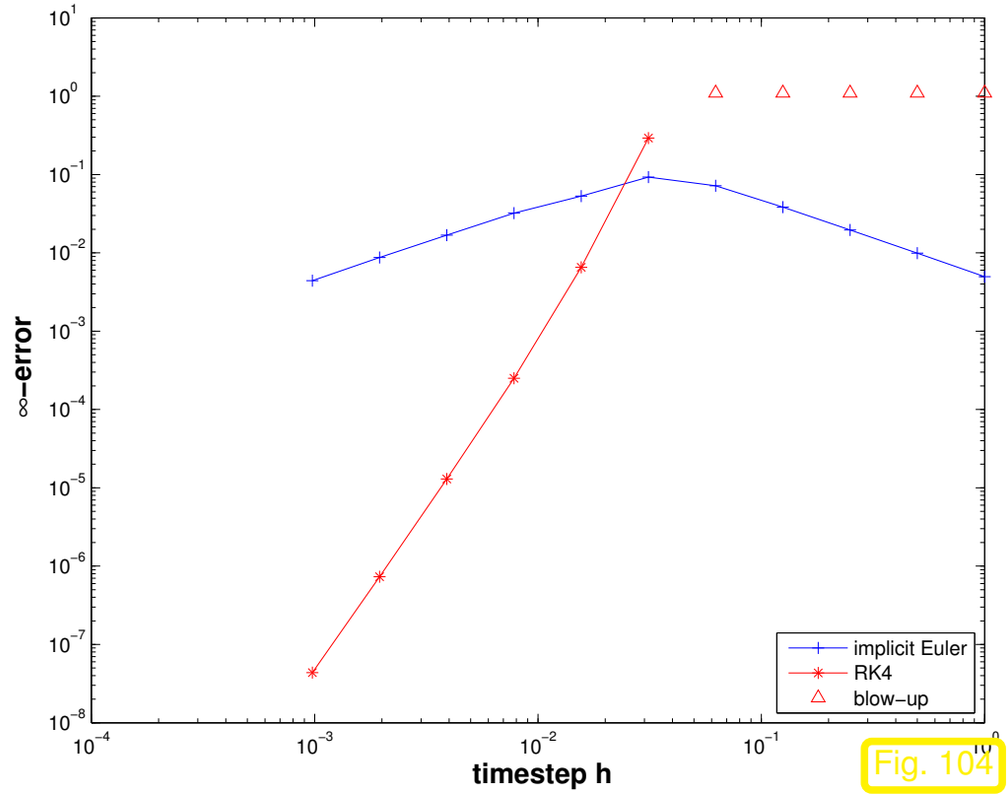
- Integrators: implicit euler-method (1.4.13), classical Runge-Kutta-method (2.3.11)
- Uniform time-stepwidth $h = 1/N$, $N \in \mathbb{N}$
- Error measure: $\mathrm{err} = \max_k |y_k - y(t_k)|$, $k = 1, \dots, N$

P. Grohs

rev 63606, February 17, 2014



(approximative) solutions for $N = 30$



errors versus stepwidth (both axes logarithmic)

P. Grohs

rev 63606, February 17, 2014

Observation:

- RK4 *asymptotic* more exact than implicit euler-method
- RK4 *preasymptotic* (for $h > 0.02$) not useful (instability)

◇

Consideration Linearised around fixpoint, see Rem. 1.3.19

3.1 Model Problem Analysis

Consideration for Ex. 3.0.1: In the neighborhood of a fixpoint the solution of a (preliminary scalar) ODE behaves as their linearization.

➢ Relevance of the (around a fixpoint) linearized ODE: Numerical integrator is only useful for a solution of the ODE near the fixpoint if he is established at least for the linearized ODE.

Linear autonomous scalar ODE are simple: $\dot{y} = \lambda y$ (upon translation)

In this section we study the behavior of numerical integrators for such simple ODEs

Nothing new! Remider to Section 1.4.1: Insights in the behavior of the explicit euler-method (1.4.2) by *model problem analysis*, i.e., analytical study of the discrete evolution for the scalar linear ODE $\dot{y} = \lambda y$, $\lambda \in \mathbb{C}$.

$$\boxed{\text{Autonomous scalar linear IVP:} \quad \dot{y} = \lambda y\,,\ y(0) = 1\,, \quad \operatorname{Re}\lambda < 0 \quad \text{auf } [0,\infty[} \tag{3.1.1}$$

$y(t) = e^{\lambda t} \to 0$ für $t \to \infty$ (so called asymptotic stability of $y = 0$) .

P. Grohs

rev 63606, February 17, 2014

Note: complex $\lambda \in \mathbb{C}$ allowed in the model problem ➢ complex state space $\mathbb{C}$
(Reason: "'diagonalize technique" for linear, autonomous IVP, Sect. 1.3.2, cf. Rem. 3.1.13)

Question: When does the solution $\{y_k\}_{k=0}^{\infty}$ "'inherit", $y_{k+1} = \Psi_\lambda^h y_k$ ($\Psi_\lambda^h \hat{=}$ discrete evolution) from RK-OSM on (infinite) equidistant grid (mesh width h) have asymptotic stability **?**

This is a question of structure-preserving: Accordance of qualitative properties of the continuum and discrete evolution.

Remark 3.1.2 (Rescaling of the model problem).

Note: Application of a linear operator on $\mathbb{R}$ $\leftrightarrow$ multiplication with a real number

$$L(\mathbb{R},\mathbb{R}) \cong \mathbb{R}$$

✎ Notation: $L(\mathbb{R},\mathbb{R}) \hat{=}$ space of linear operators on $\mathbb{R}$

Since IVP (3.1.1) autonomous & scalar

- $\Phi_\lambda^h \in L(\mathbb{R},\mathbb{R})$
- Application of Φ_λ^h on the state $y \in \mathbb{C} \sim$ multiplication
- $(h,\lambda) \mapsto \Phi_\lambda^h$ defined by the function $\mathbb{R} \times \mathbb{C} \mapsto \mathbb{C}$

Which function is this?

$$\Phi_\lambda^h(y) = e^{\lambda h} y \quad \forall y \in \mathbb{R} \quad \Rightarrow \quad \text{Function} \quad (h,\lambda) \mapsto e^{\lambda h} \,.$$

► $\boxed{\Phi_\lambda^h = \Phi_1^{\lambda h}}$ $\Rightarrow$ Function does only depend on the product λh. (3.1.3)

In addition, for the discrete evolution of a Runge-Kutta-one-step-method we have $\Psi_\lambda^h \in L(\mathbb{R},\mathbb{R})$

► $(h,\lambda) \mapsto \Psi_\lambda^h$ can also be defined by the function $\mathbb{R} \times \mathbb{C} \mapsto \mathbb{C}$

Related question: Does the time-scaling-invariance (3.1.3) also hold for Ψ_λ^h, i.e. do we have

$$\Psi_\lambda^h = \Psi_1^{\lambda h} \quad \forall \lambda \in \mathbb{C}\,,\ h \text{ sufficiently small } ?$$

P. Grohs

rev 63606, February 17, 2014

The time-scaling-invariance (3.1.3) is satisfied for Runge-Kutta-one-step-method ($\rightarrow$ Def. 2.3.5), as we can check by simple calculations! (h and λ enter the incrementequation of the RK-OSM for (3.1.1) only in the form of the product $h\lambda$, see proof of thm. 3.1.6.)

► $\boxed{\Psi_\lambda^h = \Psi_1^{\lambda h}}$ depends only on $z := \lambda h$: $S(z) := \Psi_\lambda^h$

stability function

interpreted as a number

What does this stability function tell us about the qualitative asymptotic of the discrete solution **?**

▶ discrete solution: $y_k = S(z)^k y_0 \,, \quad k \in \mathbb{N}_0 \,, \quad z := \lambda h \,.$

➥ $|S(z)| < 1 \quad \Leftrightarrow \quad \lim\limits_{k\to\infty} y_k = 0 \quad \forall y_0 \in \mathbb{R}$

$\Leftrightarrow \quad y = 0$ asymptotic stable ($\to$ Def. 3.2.2) for discrete evolution Ψ_λ^h .

Definition 3.1.4 (stability domain of a one-step-method)**.** *[12, Sect. 6.1.2]*
The stability domain of a OSM for the IVP (3.1.1) *on the basis of the discrete evolution* $\Psi_\lambda^h y =: S(z)y$, $y \in \mathbb{C}$, $z := \lambda h$, $S : D_S \subset \mathbb{C} \mapsto \mathbb{C}$, *is*

$$\mathcal{S}_\Psi := \{z \in D_S \colon |S(z)| < 1\} \subset \mathbb{C} \,.$$

☞ For the gridfunction $\{y_k\}_{k\in\mathbb{N}}$ generated by RK-OSM to IVP (3.1.1) on a equidistant mesh with mesh width $h > 0$ we have

$$y_0 \neq 0\colon \quad \lim_{k\to\infty} y_k = 0 \quad \Leftrightarrow \quad \lim_{k\to\infty} S(h\lambda)^k = 0 \quad \Leftrightarrow \quad h\lambda \in \mathcal{S}_\Psi \,. \tag{3.1.5}$$

Theorem 3.1.6 (Stabilityfunction of Runge-Kutta-method)**.**
The discrete evolution Ψ_λ^h to a s-level Runge-Kutta-one-step method ($\rightarrow$ Def. 2.3.5) with Butcher-tableau $\begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^T \end{array}$ *(see (2.3.6)) for the ODE $\dot{y} = \lambda y$ is a multiplication operator of the form*

$$\Psi_\lambda^h = \underbrace{1 + z\mathbf{b}^T\,(\mathbf{I} - z\mathfrak{A})^{-1}\,\mathbf{1}}_{\text{Stabilitätsfunktion } S(z)} = \frac{\det(\mathbf{I} - z\mathfrak{A} + z\mathbf{1}\mathbf{b}^T)}{\det(\mathbf{I} - z\mathfrak{A})}\,, \quad z := \lambda h\,, \quad \mathbf{1} = (1,\dots,1)^T \in \mathbb{R}^s\,.$$

[Beitrag von J. Mitrovic, FS 2011]

Proof: Let $\mathbf{k} = (k_1, k_2, ..., k_n)^T$ be the increment vector. The equation of the method written in matrix form yield: $\begin{pmatrix} \mathbf{I} - z\mathbf{U} & 0 \\ -z\mathbf{b^T} & \mathbf{1} \end{pmatrix} \begin{pmatrix} \mathbf{k} \\ y_1 \end{pmatrix} = \begin{pmatrix} 1 \\ y_0 \end{pmatrix}$.

solved for y_1 we get $y_1 = \underbrace{(1 + z\mathbf{b^T}(\mathbf{I} - z\mathbf{U})^{-1}\mathbf{1})}_{=\Psi_\lambda^h = S(z) = S(h\lambda)}\, y_0$, which was to be proven.

Alternative proof: From the matrix of the method solve with cramer's rule. □

Remark 3.1.7 (Interpretation of the stability function).

$$\boxed{\Psi_\lambda^h y = S(z)y = (1+\lambda h \mathbf{b}^T(\mathbf{I}-\lambda h\mathfrak{A})^{-1}\mathbf{1})y} \longleftrightarrow \boxed{\Phi_\lambda^h = e^{\lambda h}}$$

discrete evolution (continuous) evolution

➢ $S(z) \approx \exp(z)$: stability function = approximation of the exponential function (around 0).

△

Corollary 3.1.8.

explicit Runge-Kutta-method ➤ $S(z) \in \mathcal{P}_s$,

general Runge-Kutta-method ➤ $S(z) = \dfrac{P(z)}{Q(z)}$, $P, Q \in \mathcal{P}_s$.

Proof. From the determinant formula of thm. 3.1.6:

explicit Runge-Kutta-method $\Rightarrow$ $\mathfrak{A}$ proper lower triangle matrix $\Rightarrow$ $\det(\mathbf{I} - z\mathfrak{A}) = 1$

In general $z \mapsto \det(\mathbf{I} - z\mathbf{M})$, $\mathbf{M} \in \mathbb{R}^{s,s}$, is a polynomial of degree s, which follows from the combinatorial definition of the determinant. $\square$

Corollary (3.1.8) ➢ No RK-OSM can solve $\dot{y} = \lambda y$ with given stepwidth h for all $\lambda \in \mathbb{R}$ without errors, then the exponential function ($\rightarrow$ Rem. 3.1.7) can of course not be modeled by a rational function.

[Beitrag von J. Mitrovic, FS 2011]

Remark. A RK-OSM is only consistent of order p, if and only if the first p terms of the stability function S(z) coincide with the first p terms of the exp-function (resp. its taylor expansion), i.e. $\Phi_\lambda^h - \Psi_\lambda^h = O(h^{p+1}) \Rightarrow e^z - S(z) = O(h^{p+1})$ und damit $S(z) = \sum_{k=0}^{p} \frac{1}{k!} z^k + O(h^{p+1})$.

corollary (3.1.8) ➢ bound of order for explicit/implicit RK-OSM, see Sect. 2.3.2

Lemma 3.1.9 (Rational approximation of the exponential function)**.**
If $S(z) = \dfrac{P(z)}{Q(z)}$, $P, Q \in \mathcal{P}_s$, $s \in \mathbb{N}$, *we have*

$$S(z) - \exp(z) = O(|z|^m) \quad \textit{for } z \to 0 \quad \Rightarrow \quad m \leq 2s+1 \,.$$

Proof: ($\rightarrow$ [12, Lemma 6.4], but the proof there is wrong!)

Indirect Proof, Assumption $S(z) - \exp(z) = O(|z|^{2s+2})$ for $z \to 0$:

Ansatz:
$$P(z) = p_0 + p_1 z + \cdots + p_s z^s \,,$$
$$Q(z) = q_0 + q_1 z + \cdots + q_s z^s \,, \quad q_0 = 1, \text{ da O.B.d.A } Q(0) = 1 \,.$$

$$Q(z)\exp(z) - P(z) = \alpha_{2s+2} z^{2s+2} + \alpha_{2s+3} z^{2s+3} + \ldots \text{ (global convergent power series)} \,.$$

Plugging in of the exponential series and multiplication, then comparison of coefficients ➢ linear system of equations

$$\sum_{j=0}^{s} q_j \frac{1}{(i-j)!} = 0 \,, \quad i = s+1, \ldots, 2s+1 \,.$$

$$\sum_{j=0}^{s} q_j \frac{1}{(i-j)!} - p_i = 0 \,, \quad i = 0, \ldots, s \,.$$

This system has only the trivial solution, what leads to a contradiction to $q_0 = 1$. $\square$

[Beitrag von J. Mitrovic, FS 2011]

Remark. This lemma also implies, that the order of the s-level RK-OSM is bounded by $2s$.

Example 3.1.10 (stability functions of some RK-OSM).

- Explicit Euler-method (1.4.2): $\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$ $\rightarrow$ $S(z) = 1 + z$.
- Implicit Euler method (1.4.13): $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$ $\rightarrow$ $S(z) = \dfrac{1}{1-z}$.
- Explicit trapezoidal rule (2.3.3): $\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$ $\rightarrow$ $S(z) = 1 + z + \frac{1}{2}z^2$.

- Implicit midpoint rule (2.2.19): $\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$ ➢ $S(z) = \dfrac{1+\frac{1}{2}z}{1-\frac{1}{2}z}$.

- RK4-method (2.3.11): $\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array}$ ➢ $S(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4$.

◇

P. Grohs

rev 63606, February 17, 2014

Example 3.1.11 (Behavior of stability functions).

Behavior of stability functions (for real argument z):

Rem. 3.1.7 ➢ we expect that the stability functions in $z = 0$ "adapt" to $\exp(z)$, i.e., both functions coincide in the value and some lower order derivatives. The *minimum number* of the derivatives coinciding is given by the consistency order of the one-step-method.

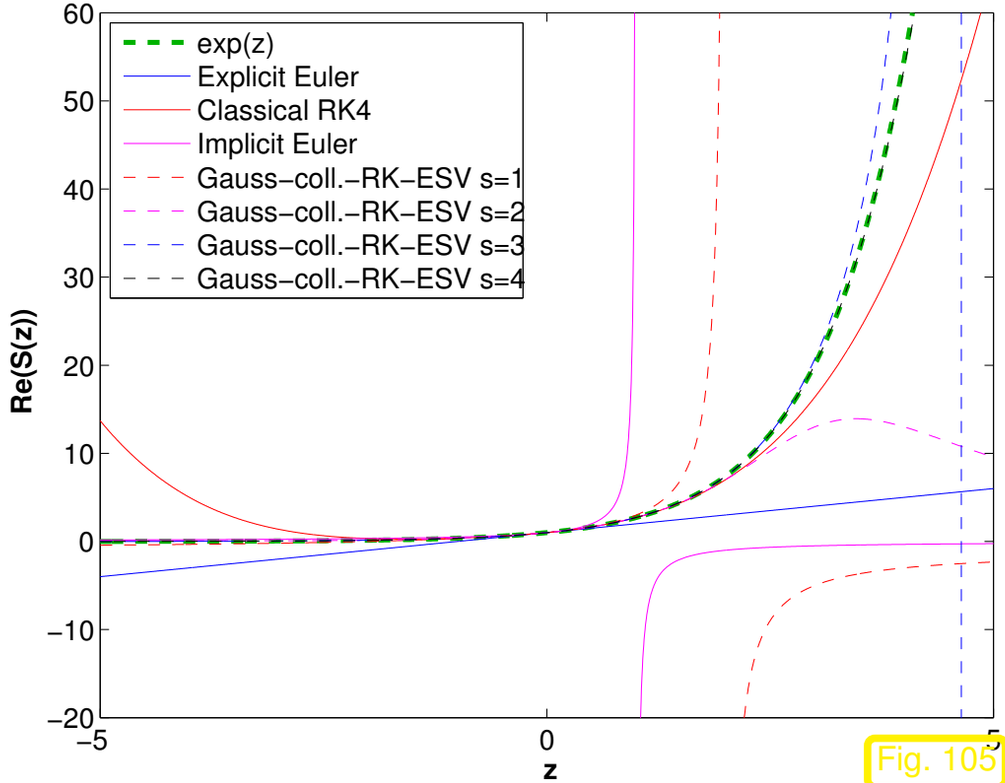


Fig. 105

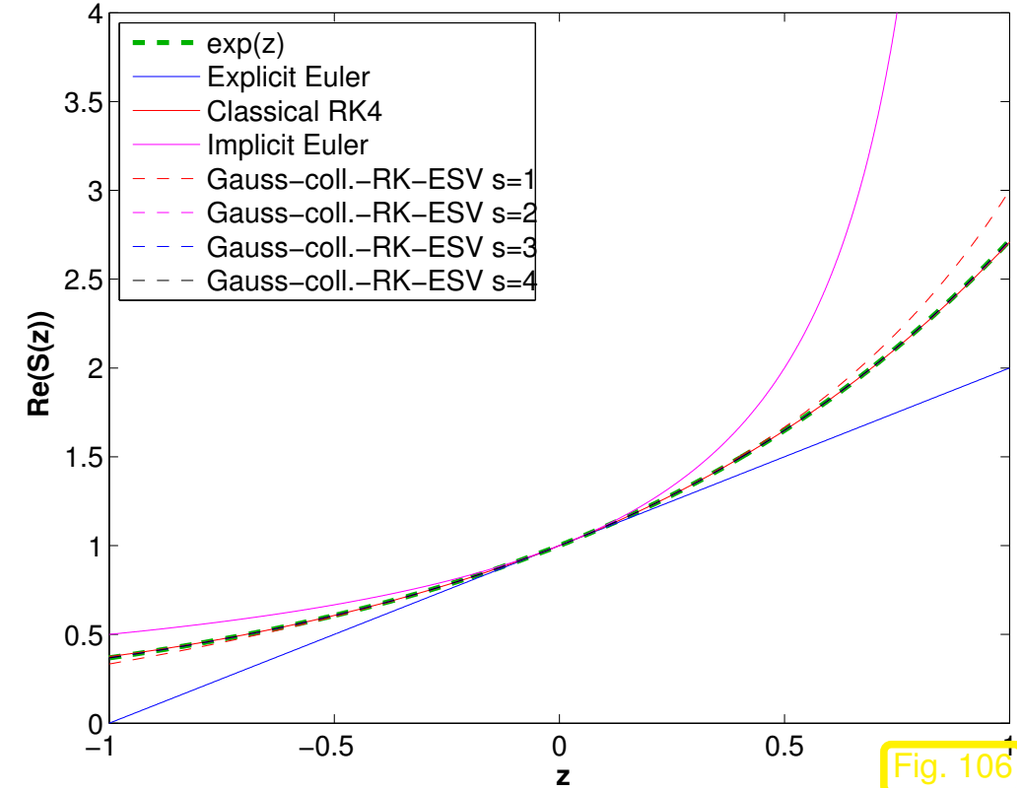


Fig. 106

◇

Examples: stability domain $\mathcal{S}$ explicit RK-OSM:

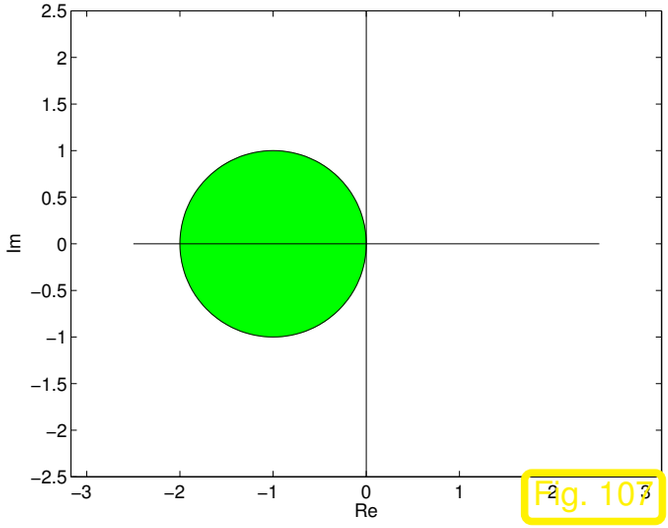


Fig. 107

$\mathcal{S}_\Psi$: expliziter Euler (2.2.1)

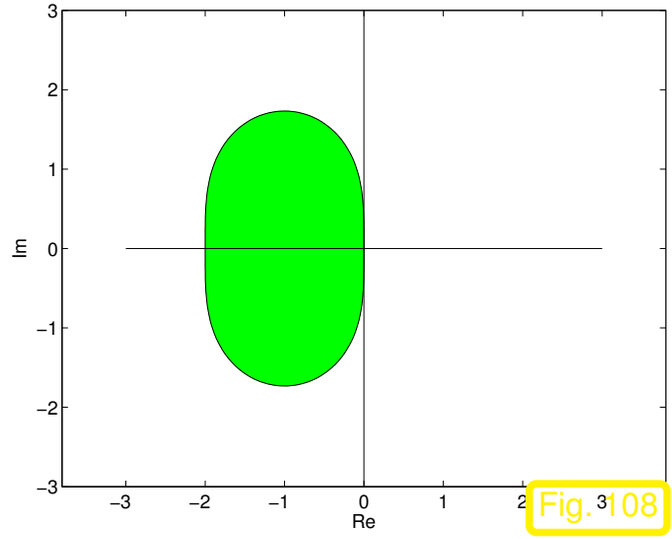


Fig. 108

$\mathcal{S}_\Psi$: explicit trapezoidal rule
midpoint rule

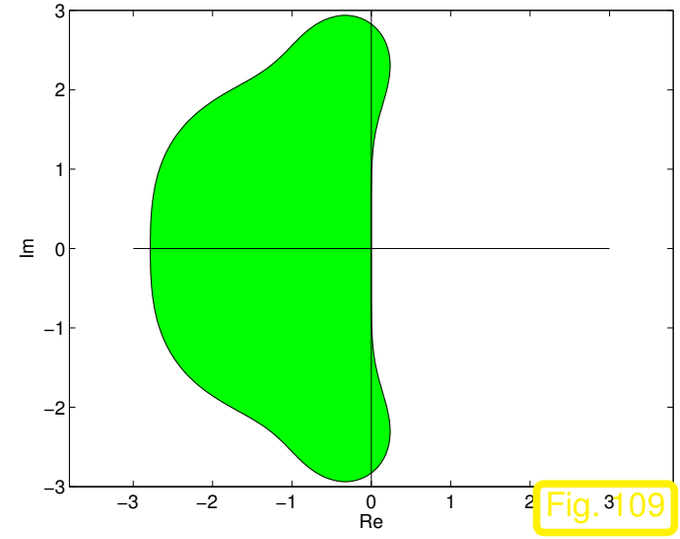


Fig. 109

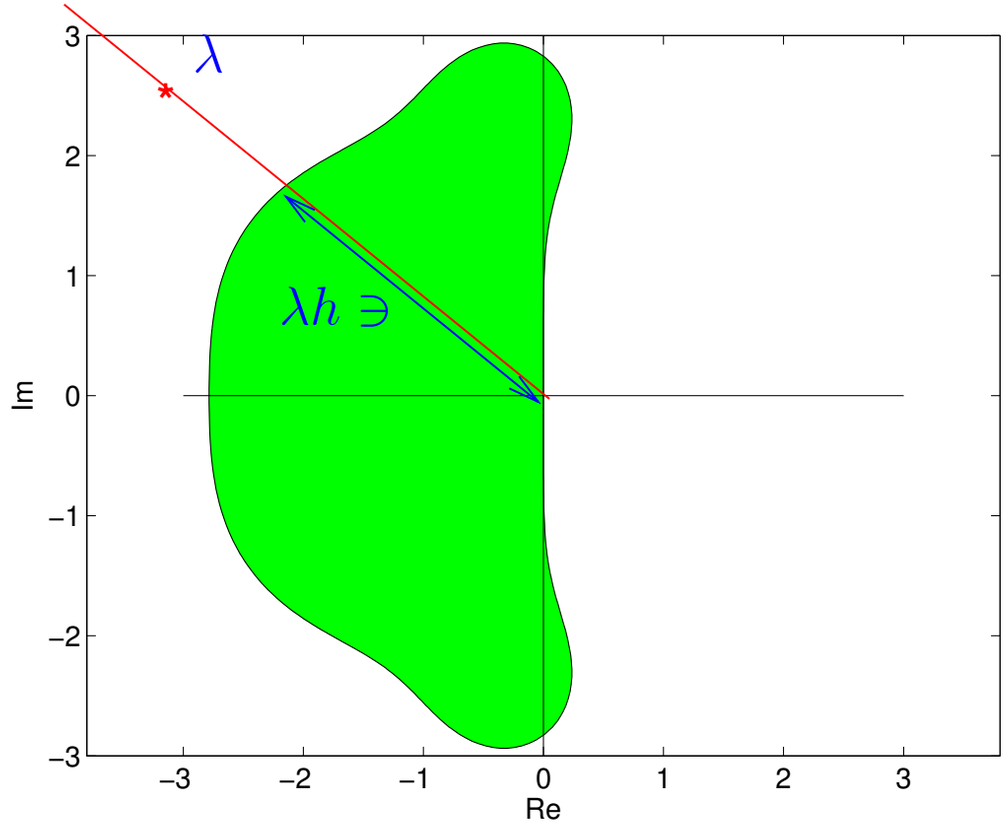
$\mathcal{S}_\Psi$: RK4-method (2.3.11)
Kuttas 3/8-rule (2.3.12)

The stability domain of explicit RK-methods is bounded!

[Beitrag von J. Mitrovic, FS 2011]

Remark. This holds because the stability function $S(z)$ is a polynomial and therefore $|z| \to \infty \Rightarrow |S(z)| \to \infty$. From this we conclude that $|S(z)| < 1$ can only hold if also $|z|$ is below a certain threshold.

Note that the order of a RK-OSM does **not** have anything to do with the size of the stability domain.



- due to stability constraint stepwidth for explicit RK-OSM applied on (3.1.1):

$$h < \sup\{t > 0 \colon t\lambda \in \mathcal{S}_{\Psi}\} \,. \qquad (3.1.12)$$

For a detailed consideration see [12, Lemma 6.6]

Remark 3.1.13 (RK-OSM for autonomous homogeneous linear ODE). see Sect. 1.3.2

What is the discrete evolution of a RK-OSM applied on

homogeneous, autonomous, linear ODE $\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}, \quad \mathbf{A} \in \mathbb{C}^{d,d}$?

① Assumption: $\mathbf{A}$ diagonalisable $\Leftrightarrow$ $\exists \mathbf{R} \in \mathbb{C}^{d,d}$ regularisable: $\mathbf{R}^{-1}\mathbf{A}\mathbf{R} = \mathbf{D} := \operatorname{diag}(\lambda_1, \dots, \lambda_d)$

Consequence from Affin-Kovarianz of RK-OSM ($\rightarrow$ Bem. 2.3.13):
Is $\widehat{\mathbf{\Psi}}$ the discrete evolution to $\frac{d}{dt}\widehat{\mathbf{y}} = \mathbf{D}\widehat{\mathbf{y}}$ (uncoupled scalar linear ODE !), then, with $\widehat{\mathbf{y}} := \mathbf{R}^{-1}\mathbf{y}$,

$$\begin{aligned}
\mathbf{\Psi}^h\mathbf{y} &\overset{(2.3.14)}{=} \mathbf{R}\widehat{\mathbf{\Psi}}^h\mathbf{R}^{-1}\mathbf{y} = \mathbf{R}\begin{pmatrix} \widehat{\Psi}^h_{\lambda_1}\widehat{y}_1 \\ \vdots \\ \widehat{\Psi}^h_{\lambda_d}\widehat{y}_d \end{pmatrix} = \mathbf{R}\begin{pmatrix} S(h\lambda_1) & & \\ & \ddots & \\ & & S(h\lambda_d) \end{pmatrix}\widehat{\mathbf{y}} \\
&= \mathbf{R}\begin{pmatrix} P(h\lambda_1) & & \\ & \ddots & \\ & & P(h\lambda_d) \end{pmatrix}\mathbf{R}^{-1}\left(\mathbf{R}\begin{pmatrix} Q(h\lambda_1) & & \\ & \ddots & \\ & & Q(h\lambda_d) \end{pmatrix}\mathbf{R}^{-1}\right)^{-1}\mathbf{y} \\
&= \mathbf{R}P(h\mathbf{D})\mathbf{R}^{-1}\left(\mathbf{R}Q(h\mathbf{D})\mathbf{R}^{-1}\right)^{-1}\mathbf{y} = P(h\mathbf{A})Q(h\mathbf{A})^{-1}\mathbf{y} = S(h\mathbf{A})\mathbf{y}\ .
\end{aligned}$$

where $S(z) = P(z)/Q(z) \hat{=}$ stability function ($\rightarrow$ thm. 3.1.6) of the RK-OSM.

② general matrix $\mathbf{A} \in \mathbb{C}^{d,d}$ with eigenvalues (counted with multiplicity) $\lambda_1, \dots, \lambda_d \in \mathbb{C}$

Tool: Schur-decomposition

Lemma 3.1.14 (Schur-decomposition). *To each matrix* $\mathbf{A} \in \mathbb{C}^{d,d}$ *there exists a* unitary matrix $\mathbf{U} \in \mathbb{C}^{d,d}$ *and a* upper triangular matrix $\mathbf{T} \in \mathbb{C}^{d,d}$ *such that*

$$\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^H \ .$$

From the Schur-decomposition for matrices follows that the diagonalizable matrices lie dense in $\mathbb{C}^{d,d}$ (rel. to the induced matrixnorm relative to the euclidean matrixnorm): add to $\mathbf{T}$ a diagonal matrix $\mathbf{D}$ with arbitrary small diagonal entries such that $\mathbf{T} + \mathbf{D}$ has pairwise different diagonal entries. Then $\mathbf{U}(\mathbf{T} + \mathbf{D})\mathbf{U}^H$ is diagonalizable, since also this matrix has pairwise different eigenvalues.

➢ If $\sigma(h\mathbf{A}) \subset D_S$,then there is also a sequence of diagonalizable matrices $\mathbf{A}_n \to \mathbf{A}$, $n \in \mathbb{N}$, $\sigma(h\mathbf{A}_n) \subset D_S$, for die which by the continuity of the matrix multiplication holds

$$P(h\mathbf{A}_n) \to P(h\mathbf{A}) \quad , \quad Q(h\mathbf{A}_n) \to Q(h\mathbf{A}) \ .$$

Because of the continuity of the matrixinversion $\mathbf{A} \mapsto \mathbf{A}^{-1}$ on $GL(d) := \{\mathbf{M} \in \mathbb{C}^{d,d}\colon \mathbf{M} \text{ regular}\}$ therefore follows

$$S(h\mathbf{A}_n) = P(h\mathbf{A}_n)Q(h\mathbf{A}_n)^{-1} \to P(h\mathbf{A})Q(h\mathbf{A})^{-1} = S(h\mathbf{A}) \ . \tag{3.1.15}$$

If $\mathbf{\Psi}_n^h$ is the discrete evolution to $\dot{\mathbf{y}} = \mathbf{A}_n\mathbf{y}$, we have

$$\mathbf{\Psi}^h\mathbf{y} = \lim_{n\to\infty} \mathbf{\Psi}_n^h\mathbf{y} \overset{①}{=} \lim_{n\to\infty} S(h\mathbf{A}_n)\mathbf{y} \overset{(3.1.15)}{=} S(h\mathbf{A}) \ .$$

Stability function

$$\mathbf{\Psi}^h\mathbf{y} = S(h\mathbf{A})\mathbf{y} \quad \forall \mathbf{y} \in \mathbb{C}^d \ . \tag{3.1.16}$$

△

P. Grohs

rev 63606, February 17, 2014

Remark 3.1.17 (functional calculus for matrices).

For $\mathbf{A} \in \mathbb{R}^{d,d}$:

- It is clear that $p(\mathbf{A}) = \sum_{j=1}^{s} c_j \mathbf{A}^j$ for polynomial $p \in \mathcal{P}_s$, $p(z) = \sum_{j=1}^{s} c_j z^j$.

- For rational function

$$R(z) = \frac{\sum_{j=1}^{s} p_j z^j}{\sum_{j=1}^{s} q_j z^j} \quad \blacktriangleright \quad R(\mathbf{A}) = \left(\sum_{j=1}^{s} q_j \mathbf{A}^j\right)^{-1} \left(\sum_{j=1}^{s} p_j \mathbf{A}^j\right) , \tag{3.1.18}$$

if $\sum_{j=1}^{s} q_j \mathbf{A}^j$ invertible.

- If $f(z) = \sum_{i=0}^{\infty} a_j z^j$ is a power series with convergence radius $\rho > 0$, then we have

$$f(\mathbf{A}) := \sum_{i=0}^{\infty} a_j \mathbf{A}^j \quad \text{well-defined for } \|\mathbf{A}\| < \rho \ .$$

Therefore transcendental functions of matrices as the matrixexponentialfunction (1.3.14) can be defined.

For all matrixfunkcions introduced above we have cf. 1.3.15,

$$\boxed{\mathbf{A} = \mathbf{S}^{-1}\mathbf{B}\mathbf{S} \quad \Rightarrow \quad f(\mathbf{A}) = \mathbf{S}^{-1} f(\mathbf{B})\mathbf{S}} \quad \forall \mathbf{A}, \mathbf{B} \in \mathbb{C}^{d,d}, \quad \mathbf{S} \in \mathbb{C}^{d,d} \text{ regular} . \tag{3.1.19}$$

For the spectra we have

$$\boxed{\sigma(f(\mathbf{A})) = f(\sigma(\mathbf{A})) := \{f(\lambda) \colon \lambda \in \sigma(\mathbf{A})\} .} \tag{3.1.20}$$

3.2 Inheritance of asymptotic Stability

3.2.1 Attractive fixpoints

Consider: autonomous IVP $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\mathbf{f} \in C^1(D, \mathbb{R}^d)$, $D \subset \mathbb{R}^d$ open.

Definition 3.2.1 (fixpoint)**.**

$\mathbf{y}^*$ *is fixpoint (stationary point) of* $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$*, if* $\mathbf{f}(\mathbf{y}^*) = 0$.

The conception is clear: a fixpoint represents a state, which does not change during the evolution:

$$\mathbf{y}(0) = \mathbf{y}^* \quad \Rightarrow \quad \mathbf{y}(t) = \mathbf{y}^* \quad \forall t \in \mathbb{R}\,.$$

Definition 3.2.2 (Asymptotic stability of a fixpoint). $\rightarrow$ *[12, Def. 3.19]*

Fixpoint $\mathbf{y}^* \in D$ *asymptotic stable (attractive)*

$$:\Leftrightarrow \qquad \exists \delta > 0\colon \quad \|\mathbf{y}_0 - \mathbf{y}^*\| < \delta \quad \Rightarrow \quad \mathbb{R}_0^+ \subset J(\mathbf{y}_0) \quad \wedge \quad \lim_{t\to\infty} \mathbf{y}(t) = \mathbf{y}^*\,,$$

where $\mathbf{y}(t)$ *is the solution of the IVP* $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\mathbf{y}(0) = \mathbf{y}_0$.

P. Grohs

rev 63606, February 17, 2014

Reminder to Def. 1.3.1: $J(\mathbf{y}_0) \hat{=}$ maximal existence interval of the solution of a autonomous differential equation to the initial value $\mathbf{y}_0$.

"Asymptotic stability" in words: A fixpoint state $\mathbf{y}^*$ is asymptotic stable/attractive, if all solution curves, which start sufficiently close to it converge to $\mathbf{y}^*$.

The following example gives a visual imagination:

Example 3.2.3 (Attractive and repulsive fixpoints of a scalar ODE).

Solution curves of the ODE ▷

$$\dot{y} = -y(1-y)(1+y)$$

$y^* = 0$ is asymptotic stable (attractive) fixpoint,
$y^* = \pm 1$ are instable (repulsive) fixpoints

(→ Ex. 1.2.1)

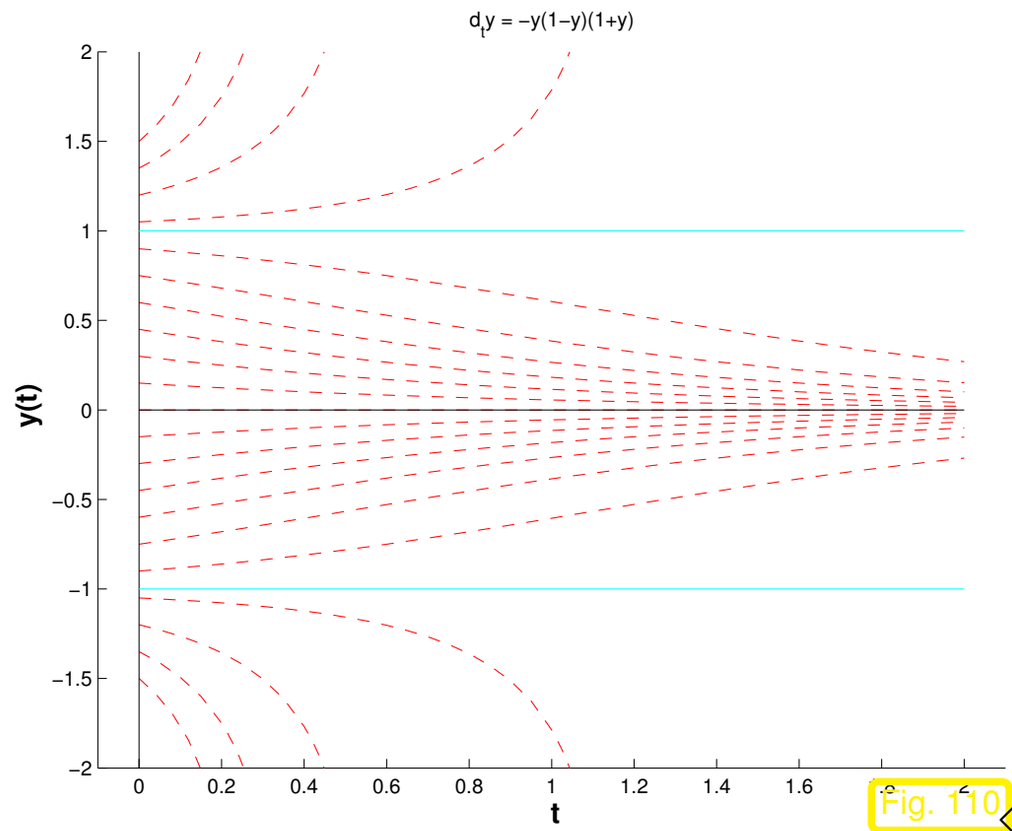


Fig. 110 ◇

Theorem 3.2.4 (Sufficient condition for asymptotic stability)**.**
fixpoint $\mathbf{y}^* \in D$ *is asymptotic stable, if*

$$\sigma(D\mathbf{f}(\mathbf{y}^*)) \subset \mathbb{C}^- := \{z \in \mathbb{C}\colon \operatorname{Re} z < 0\} \ .$$

✎ Notation: $\sigma(\mathbf{A}) := \{\lambda : \ \lambda \text{ ist Eigenwert von } \mathbf{A}\} \hat{=}$ Spectrum of a matrix

Tool: Matrixexponentialfunction (1.3.14), Jordan-Normalform of a $\mathbf{A} \in \mathbb{C}^{d,d}$:

$$\exists \mathbf{S} \in \mathbb{C}^{d,d} \text{ regular: } \quad \mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \operatorname{diag}(\mathbf{J}_1, \dots, \mathbf{J}_m) \ ,$$

with Jordan-blocs of the form ($\lambda \in \sigma(\mathbf{A})$)

$$\mathbf{J}_k = \begin{pmatrix} \lambda & 1 & 0 & \dots & \dots & 0 \\ 0 & \lambda & 1 & 0 & & \vdots \\ & & \ddots & \ddots & & \\ & & & & \lambda & 1 \\ & & & & & \lambda \end{pmatrix} = \lambda \mathbf{I} + \mathbf{N}_k \in \mathbb{C}^{d_k,d_k}, \quad d_k \in \{1, \dots, d\} \ .$$

$\mathbf{N}_k \in \mathbb{C}^{d_k,d_k}$ are nilpotent matrices: $\quad \mathbf{N}^{d_k} = 0$

Because of (1.3.15) it is enough to study $\exp(\mathbf{J})$ for a generic Jordan-bloc $\mathbf{J} \in \mathbb{C}^{n,n}$.

Use: $\quad \mathbf{A}, \mathbf{B} \in \mathbb{R}^{n,n}$: $\quad \mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A} \quad \Rightarrow \quad \exp(\mathbf{A} + \mathbf{B}) = \exp(\mathbf{A}) \cdot \exp(\mathbf{B}) \ .$ (3.2.5)

$$\blacktriangleright \quad \exp(t\mathbf{J}) = \exp(t\lambda\mathbf{I} + \lambda\mathbf{N}_k) = \exp(t\lambda\mathbf{I}) \exp(t\mathbf{N}_k) = e^{\lambda t} \exp(t\mathbf{N}_k) \ .$$

Note: $\exp(t\mathbf{N})$ is a *polynomial* in t if $\mathbf{N}$ nilpotent (exponential series ends after finitely many terms). Hence we find

$$\exp(\mathbf{A}t) = \mathbf{S}\exp(\mathbf{D}t)\mathbf{P}(t)\mathbf{S}^{-1}\ , \quad \mathbf{D} = \operatorname{diag}(\lambda_1,\dots,\lambda_d)\ ,$$

with a matrix polynomial $\mathbf{P}$ of degree $< d$, $\lambda_1,\dots,\lambda_d \hat{=}$ eigenvalues of $\mathbf{A}$ (with multiplicity).

$$\blacktriangleright \qquad \|\exp(\mathbf{A}t)\| \le \|\mathbf{S}\| \left\|\mathbf{S}^{-1}\right\| \|\exp(\mathbf{D}t)\| \cdot \|\text{matrixpolynomial in } t\|\ ,$$

where $\mathbf{D}$ is the diagonalmatrix of the eigenvalues of $\mathbf{A}$. Because of

$$\forall\lambda\in\mathbb{R}:\ \ \forall\beta>\lambda:\ \ \forall p\in\mathcal{P}_n:\ \ \exists C = C(\lambda,\beta,p):\ \ e^{\lambda t}p(t) \le Ce^{\beta t} \quad \forall t\in\mathbb{R}\ .$$

we conclude: $\|\exp(\mathbf{A}t)\| \le Ce^{\beta t}$ for each $\beta > \max\{\operatorname{Re}\sigma(\mathbf{A})\}$.

P. Grohs

rev 63606, February 17, 2014

Proof. (of Thm.3.2.4) $\to$ [12, Satz 3.30], Wlog $\mathbf{y}^* = 0$

Linearized, see Rem. 1.3.19:

$$\mathbf{f}(\mathbf{y}) = D\mathbf{f}(0)\mathbf{y} + r(\mathbf{y})\ , \quad \|r(\mathbf{y})\| = o(\|\mathbf{y}\|) \quad \text{für } \mathbf{y}\to 0\ .$$

$\mathbf{y}(t) \hat{=}$ solution of the IVP to the initial value $\mathbf{y}_0$, $t \in J(\mathbf{y}_0)$. Variation of constant-formula, see Sect. 1.3.2:

$$\mathbf{y}(t) = \exp(D\mathbf{f}(0)t)\mathbf{y}_0 + \int_0^t \exp(D\mathbf{f}(0)(t-\tau))r(\mathbf{y}(\tau))\,\mathrm{d}\tau\ . \tag{3.2.6}$$

With the help of the Jordan-normal form, see above:

$$\forall \beta \in]\underbrace{\max\{\operatorname{Re}\lambda : \ \lambda \in \sigma(D\mathbf{f}(0))\}}_{<0\,!}, 0[: \quad \exists C = C(\beta) > 0: \quad \|\exp(D\mathbf{f}(0)t)\| \leq Ce^{\beta t} \quad \forall t \in \mathbb{R}\,.$$

Fix an appropriate $\beta < 0$ and $C > 0$. In addition there is an $\epsilon > 0$: $\|r(\mathbf{y})\| \leq \frac{|\beta|}{2C}\|\mathbf{y}\|$,if $\|\mathbf{y}\| < \epsilon$

Assumption: $\|\mathbf{y}(t)\| < \epsilon$ for $0 < t < \delta$. So that for $0 \leq t < \delta$ from (3.2.6)

$$\|\mathbf{y}(t)\| \leq Ce^{\beta t}\|\mathbf{y}_0\| + \frac{|\beta|}{2}\int_0^t e^{\beta(t-\tau)}\|\mathbf{y}(\tau)\|\,\mathrm{d}\tau\,,$$

$$\blacktriangleright \quad e^{|\beta|t}\|\mathbf{y}(t)\| \leq C\|\mathbf{y}_0\| + \frac{|\beta|}{2}\int_0^t e^{|\beta|\tau}\|\mathbf{y}(\tau)\|\,\mathrm{d}\tau$$

Use: Gronwalls Lemma (Lemma 1.3.29) for $u(t) := e^{|\beta|t}\|\mathbf{y}(t)\|$

$$\blacktriangleright \quad \|\mathbf{y}(t)\| \leq C\|\mathbf{y}_0\|\exp(-\tfrac{|\beta|}{2}t) \quad \forall 0 \leq t < \delta\,. \tag{3.2.7}$$

Now we see, that the assumption $\|\mathbf{y}(t)\| < \epsilon$ is satisfied, if $\|\mathbf{y}_0\| < \dfrac{\epsilon}{\max\{C,1\}}$.

Under this condition we have (3.2.7) for all $t \geq 0$ and also $\mathbb{R}_0^+ \subset J(\mathbf{y}_0)$ with Thm. 1.3.4. □

Asymptotic stability of a fixpoint $\mathbf{y}^*$ follows from the asymptotic stability of the fixpoints $\mathbf{y}^*$ of the around $\mathbf{y}^*$ linearised ODE

$$\dot{\mathbf{y}} = D\mathbf{f}(\mathbf{y}^*)(\mathbf{y} - \mathbf{y}^*) \; . \tag{3.2.8}$$

This confirms the intuition motivated by the model problem analysis of Sect. 3.1 that the behavior of solutions of an ODE in a environment of a fixpoint is described correctly by the behavior of the solutions around the linearized ODE.

3.2.2 Attractive Fixpoints of One-Step-Methods

We still consider an autonomous IVP $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\mathbf{f} \in C^1(D, \mathbb{R}^d)$, $D \subset \mathbb{R}^d$ open.

Furthermore let $\mathbf{y}^*$ be a fixpoint ($\rightarrow$ Def. 3.2.1): $\mathbf{f}(\mathbf{y}^*) = 0$.

Consider: (consistent) RK-OSM for autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ ($\rightarrow$ Def. 2.3.5)

$$\mathbf{k}_i := \mathbf{f}(\mathbf{y} + h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j)\ , \quad i = 1,\dots,s \quad , \quad \mathbf{\Psi}^h\mathbf{y} := \mathbf{y} + h\sum_{i=1}^{s} b_i\mathbf{k}_i\ . \tag{3.2.9}$$

Sufficient: $\sum_{i=1}^{s} b_i = 1$, Lemma 2.3.23

Assumption: h sufficiently small for well-defined of the OSM for $\mathbf{y}$ "'close to" $\mathbf{y}^*$, $\rightarrow$ Lemma. 2.2.7.

P. Grohs

rev 63606, February 17, 2014

$$\blacktriangleright \quad \mathbf{\Psi}^h\mathbf{y}^* = \mathbf{y}^* \quad \forall h \text{ sufficiently small}\ . \tag{3.2.10}$$

Consider with the help of the map $\Pi : D \subset \mathbb{R}^d \mapsto \mathbb{R}^d$ recursively defined sequence $\mathbf{y}_{k+1} = \Pi(\mathbf{y}_k)$.

(We also say that Π is a discrete dynamical system defined. Obviously all One-step-methods represent a discrete dynamical system.)

Clear: $\mathbf{y}^* \in D$ is called fixpoint of the discrete dynamical system, if $\Pi(\mathbf{y}^*) = \mathbf{y}^*$.

Also clear: definitions of the asymptotic stability of a fixpoint of a discrete dynamical system analogously to Def. 3.2.2

Theorem 3.2.12 (Asymptotic stability of fixpoints of discrete dynamical systems)**.**
Let $\Pi : D \subset \mathbb{R}^d \mapsto \mathbb{R}^d$ be continuously differentiable and $\Pi(\mathbf{y}^*) = \mathbf{y}^*$ for a $\mathbf{y}^* \in D$. Then we have

$$\rho(D\Pi(\mathbf{y}^*)) < 1 \quad \Rightarrow \quad \mathbf{y}^* \text{ is asymptotically stable fixpoint of } \mathbf{y}_{k+1} := \Pi(\mathbf{y}_k) \ .$$

✎ Notation: $\rho(\mathbf{A}) := \max\{|\lambda| : \ \lambda \in \sigma(\mathbf{A})\} \hat{=}$ spectral radius of a matrix

Lemma 3.2.13 (The spectral radius of approximated matrix norm). $\rightarrow$ *[16, Sect. 2.9.3]*
To each matrix $\mathbf{A} \in \mathbb{C}^{d,d}$ and each $\epsilon > 0$ there is a vector norm $\|\cdot\|_{A,\epsilon}$ on $\mathbb{R}^d$ such that for the induced matrix norm we have

$$\rho(\mathbf{A}) \leq \|\mathbf{A}\|_{A,\epsilon} \leq \rho(\mathbf{A}) + \epsilon \ .$$

The proof is based on the schur-normalform of $\mathbf{A}$.

[Beitrag von J. Mitrovic, FS 2011]

Proof of Thm. 3.2.12. Since $\rho(\mathbf{M}) < 1$ for matrix $\Rightarrow \exists \, ||.||$ norm : $||\mathbf{M}|| < 1$. $\Rightarrow ||y_{k+1} - y^\star|| = ||\Pi y_k - \Pi y^\star|| \leq ||D\Pi(y^\star)(y_k - y^\star)|| + ||r(y_k - y^\star)||$, where we put transformation with the triangle inequality and the taylor expansion around $y^\star$ in execution.

From $r(y) = o(||y||)$ follows: $\exists \epsilon > 0 : \ ||y|| \leq \epsilon \Rightarrow ||r(y)|| \leq \frac{1}{2}(\mathbf{I} - ||D\Pi(y^\star)||) \, ||y||$.

With $||y_k - y^\star|| \leq (||y_k - y^\star)||) \underbrace{(||D\Pi(y^\star)|| + \frac{1}{2}(\mathbf{I} - ||D\Pi(y^\star)||))}_{\underbrace{\frac{1}{2}(\mathbf{I} + ||D\Pi(y^\star)||}_{<1}}$ follows $y_k \rightarrow y^\star$ für $k \rightarrow$

∞. $\square$

☞ For discrete evolution $\mathbf{\Psi}^h$: Study the Jacobi-matrix $D_{\mathbf{y}}(\mathbf{\Psi}^h\mathbf{y})$ for $\mathbf{y} = \mathbf{y}^*$!

Für RK-OSM: $$\boxed{D_{\mathbf{y}}(\mathbf{\Psi}^h)(\mathbf{y}^*) = S(hD\mathbf{f}(\mathbf{y}^*))} \,. \tag{3.2.14}$$

Reminder: S is the (rationale) stability function ($\rightarrow$ Thm. 3.1.6) of the RK-OSM, in (3.2.14) used in the sense of Rem. 3.1.17.

$$\{\mathbf{f}(\mathbf{y}^*) = 0 \quad \Leftrightarrow \quad \mathbf{\Phi}^h\mathbf{y}^* = \mathbf{y}^*\} \qquad \Rightarrow \qquad \mathbf{\Psi}^h\mathbf{y}^* = \mathbf{y}^*$$

$$\Downarrow$$

$$\mathbf{y}(h) \approx (\mathbf{y}_0 - \mathbf{y}^*)\exp(D\mathbf{f}(\mathbf{y}^*)h) + \mathbf{y}^* \;\leftrightarrow\; \mathbf{\Psi}^h\mathbf{y} \approx (\mathbf{y}_0 - \mathbf{y}^*)S(hD\mathbf{f}(\mathbf{y}^*)) + \mathbf{y}^*$$

[Beitrag von J. Mitrovic, FS 2011]

Derivation of the equation above: We have

$$D_y k_i = D_y f(\mathbf{y} + h \sum_{i=1}^{s} a_{ij} k_j) = Df(\mathbf{y} + h \sum_{i=1}^{s} a_{ij} k_j)(\mathbf{I} + h \sum_{i=1}^{s} a_{ij} D_y k_j)$$

and

$$D_y \Psi^h \mathbf{y} = \mathbf{I} + h \sum_{i=1}^{s} b_i D_y k_i \ .$$

For $\mathbf{y} = \mathbf{y}^\star \Rightarrow k_i = 0$.

From this we get

$$D_y k_i(\mathbf{y}^\star) = Df(\mathbf{y}^\star)(\mathbf{I} + h \sum_{i=1}^{s} a_{ij} D_y k_j(\mathbf{y}^\star)) \ ,$$

and

$$D_y \Psi^h(y^\star) = \mathbf{I} + h \sum_{i=1}^{s} b_i D_y k_i(y^\star) \ .$$

Thought about RK-OSM applied on matrix-ODE:

$$\dot{\mathbf{y}} = A\mathbf{y} \quad \mathbf{y}(0) = \mathbf{I} \Rightarrow k_i = A(\mathbf{I} + h\sum_{j=1}^{s} a_{ij}k_j) \quad y_1 = \mathbf{I} + h\sum_{1=1}^{s} b_i k_i$$

The matrix-ODE is homogeneous and linear $\Rightarrow y_1 = S(hA)y_0 = S(hA)\mathbf{I}$.

With comparison we get the desired. □

Theorem 3.2.15 (Inheritance of asymptotic stability)**.**
A fixpoint $\mathbf{y}^* \in D$ of the discrete evolution of a RK-OSM with stability domain $\mathcal{S}_{\mathbf{\Psi}}$ is asymptotically stable ($\rightarrow$ Def. 3.2.2) if

$$h\sigma(D\mathbf{f}(\mathbf{y}^*)) \subset \mathcal{S}_{\mathbf{\Psi}} \ .$$

[Beitrag von J. Mitrovic, FS 2011]

Proof: $y^\star$ asymptotically stable $\Rightarrow \mathbf{Re}(\sigma(D\mathbf{f}(\mathbf{y}^\star))) \leq 0$;

$h\sigma(D\mathbf{f}(\mathbf{y}^\star)) \in S_\Psi \Rightarrow |S(h\lambda)| \leq 1 \quad \forall \lambda \in \sigma(D\mathbf{f}(\mathbf{y}^\star)) \Rightarrow \rho(\underbrace{D\Psi^h(\mathbf{y}^\star)}_{S(hD\mathbf{f}(y^\star))}) = max\{|S(h\lambda)|; \lambda \in \sigma(D\mathbf{f}(\mathbf{y}^\star))\} \leq 1.$

According to the last theorem it follows that $y^\star$ is an asymptotically stable fixpoint.

Explicit RK-OSM : stepwidth constraint for inheritance of stability of a fixpoint ($\rightarrow$ (3.1.12))

For which methods is stepwidth constraint dispensed **?**

Definition 3.2.16 (A-stable one-step-method)**.** $\rightarrow$ *[12, Sect. 6.1.3]*

OSM A-stable $:\Leftrightarrow$ $\mathbb{C}^- := \{z \in \mathbb{C}\colon \operatorname{Re} z < 0\} \subset \mathcal{S}_\Psi$ *($\mathcal{S}_\Psi$ = stability domain, $\rightarrow$ Def. 3.1.4)*

[Beitrag von J. Mitrovic, FS 2011]

Remark. A-stability is valid only for implicit RK-OSM, because explicit RK-OSM have bounded stability domains.

From Thm. 3.2.15: for A-stable OSM (with discrete evolution $\mathbf{\Psi}^h$)

autonomous differential equation $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\wedge$ $\sigma(D\mathbf{f}(\mathbf{y}^*)) \subset \mathbb{C}^-$ $\Rightarrow$ If $\|\mathbf{y}_0 - \mathbf{y}^*$

fixpoint $\mathbf{f}(\mathbf{y}^*) = 0$ $(\hat{=}$ asymp. stability of $\mathbf{y}^*)$ then $\lim_{k\to\infty} \left(\mathbf{\Psi}^h\right.$

Example 3.2.17 (Simple A-stable RK-OSM).

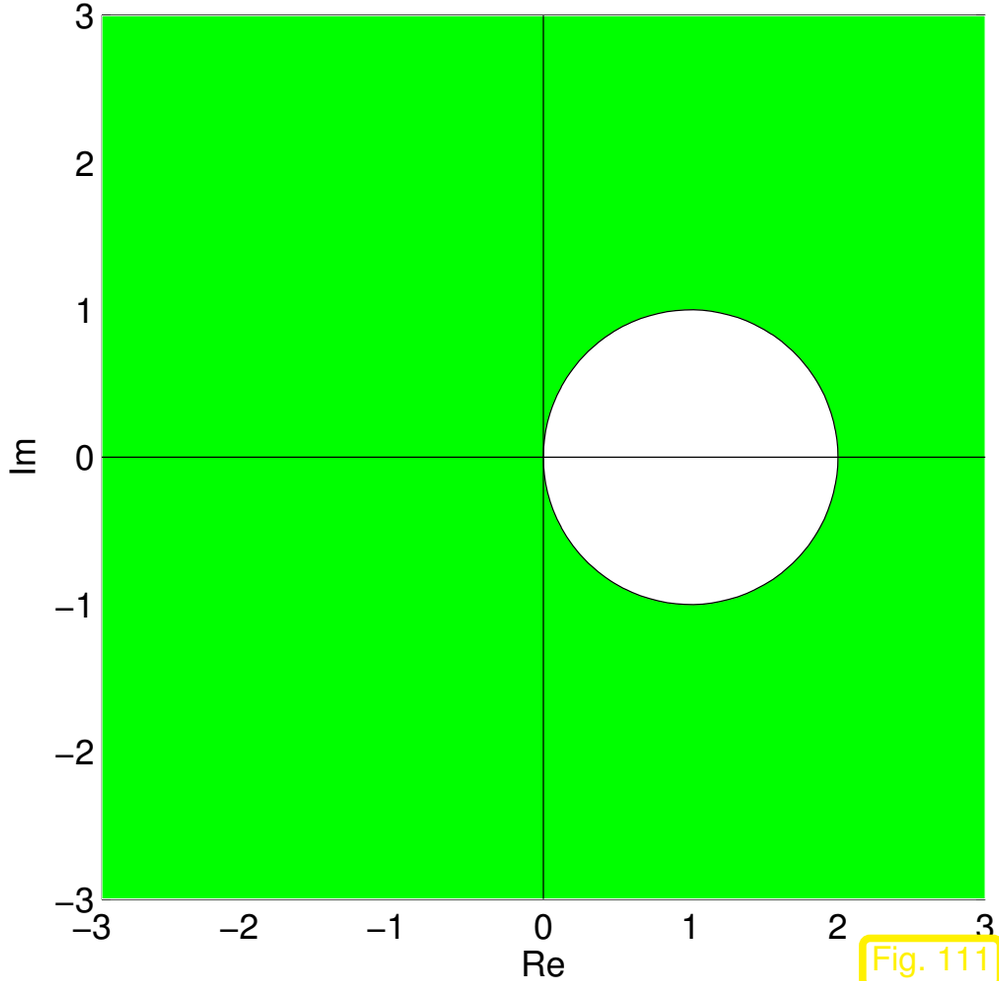


Fig. 111

Implicit euler-method (1.4.13)

stability function ($\longrightarrow$ Thm. 3.1.6)

$$S(z) = \frac{1}{1-z} \; .$$

$\triangleleft$ stability domain $\mathcal{S}_{\Psi}$ ($\longrightarrow$ Def. 3.1.4)

P. Grohs

rev 63606, February 17, 2014

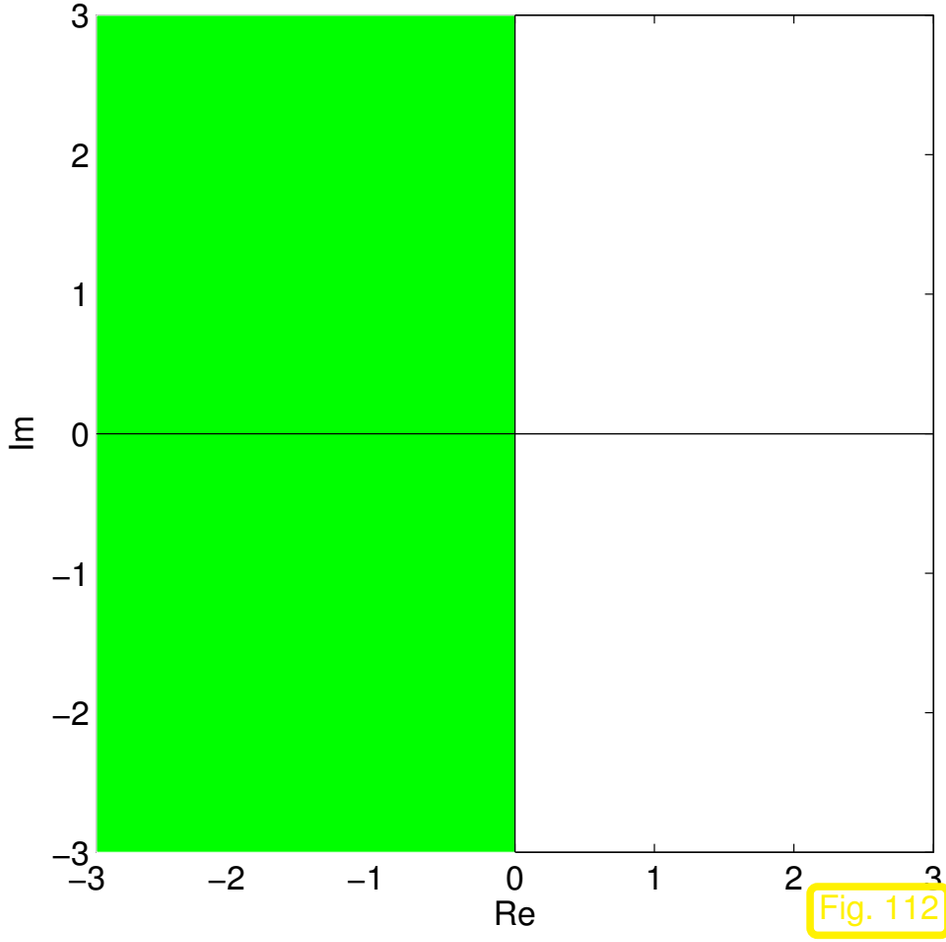


Fig. 112

implicit midpoint rule (1.4.19)

stability function ($\longrightarrow$ Thm. 3.1.6)

$$S(z) = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z} \,.$$

$\lhd$ stability domain $\mathcal{S}_{\boldsymbol{\Psi}}$ ($\longrightarrow$ Def. 3.1.4)

This is the “ideal stability domain”!

$\diamond$

3.3 Non-expansivity [12, Abschn. 6.3.3]

Consider: Autonomous IVP $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\mathbf{f} \in C^1(D, \mathbb{R}^d)$, $D \subset \mathbb{R}^d$ open.

We fix $\mathbf{M} \in \mathbb{R}^{d,d}$ s.p.d. ➢ norm $\|\mathbf{y}\|_M := (\mathbf{y}^T\mathbf{M}\mathbf{y})^{1/2}$ on $\mathbb{R}^d$.

Definition 3.3.1 (Non-expansivity).
An evolution $\mathbf{\Phi}^t$ to an autonomous differential equation resp. a discrete evolution $\mathbf{\Psi}^h$ to a corresponding one-step-method is called non-expansiv, if

$$\left.\begin{aligned} \left\|\mathbf{\Phi}^t\mathbf{y} - \mathbf{\Phi}^t\widetilde{\mathbf{y}}\right\|_M &\leq \|\mathbf{y} - \widetilde{\mathbf{y}}\|_M \ , \\ \left\|\mathbf{\Psi}^h\mathbf{y} - \mathbf{\Psi}^h\widetilde{\mathbf{y}}\right\|_M &\leq \|\mathbf{y} - \widetilde{\mathbf{y}}\|_M \end{aligned}\right. \quad \forall \mathbf{y}, \widetilde{\mathbf{y}} \in D \ ,$$

and for all $t \in J(\mathbf{y}) \cap J(\widetilde{\mathbf{y}}) \cap \mathbb{R}_0^+$ and all "'sufficiently small" $h > 0$.

Example 3.3.2 (Gradient-flow ➞ "'crawling-procedure").

Given: C^1-potential $V : \mathbb{R}^d \mapsto \mathbb{R}$ konvex

P. Grohs

rev 63606, February 17, 2014

Reminder: A map $V : \mathbb{R}^d \mapsto \mathbb{R}$ is called convex, if

$$V(\xi\mathbf{x} + (1-\xi)\mathbf{y}) \leq \xi V(\mathbf{x}) + (1-\xi)V(\mathbf{y}) \quad \forall 0 \leq \xi \leq 1 \,. \tag{3.3.3}$$

Reminder: A C^1-function $\varphi : \mathbb{R} \mapsto \mathbb{R}$ is convex if and only if φ' is monotonically increasing.

Obvious consequence from (3.3.3): If $V : \mathbb{R}^d \mapsto \mathbb{R}$ is convex then this holds true for every "cut" $\tau \mapsto V(\mathbf{y} + \tau(\mathbf{x} - \mathbf{y}))$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

$$\Downarrow$$

$$C^1\text{-potential } V : \mathbb{R}^d \mapsto \mathbb{R} \text{ convex}$$

$$\Updownarrow$$

$$\frac{d}{d\tau}\varphi \text{ monotonically increasing for } \varphi(\tau) := V(\mathbf{y} + \tau(\mathbf{x} - \mathbf{y})) \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$$

$$\Updownarrow$$

$$(\mathbf{grad}\, V(\mathbf{x}) - \mathbf{grad}\, V(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) \geq 0 \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

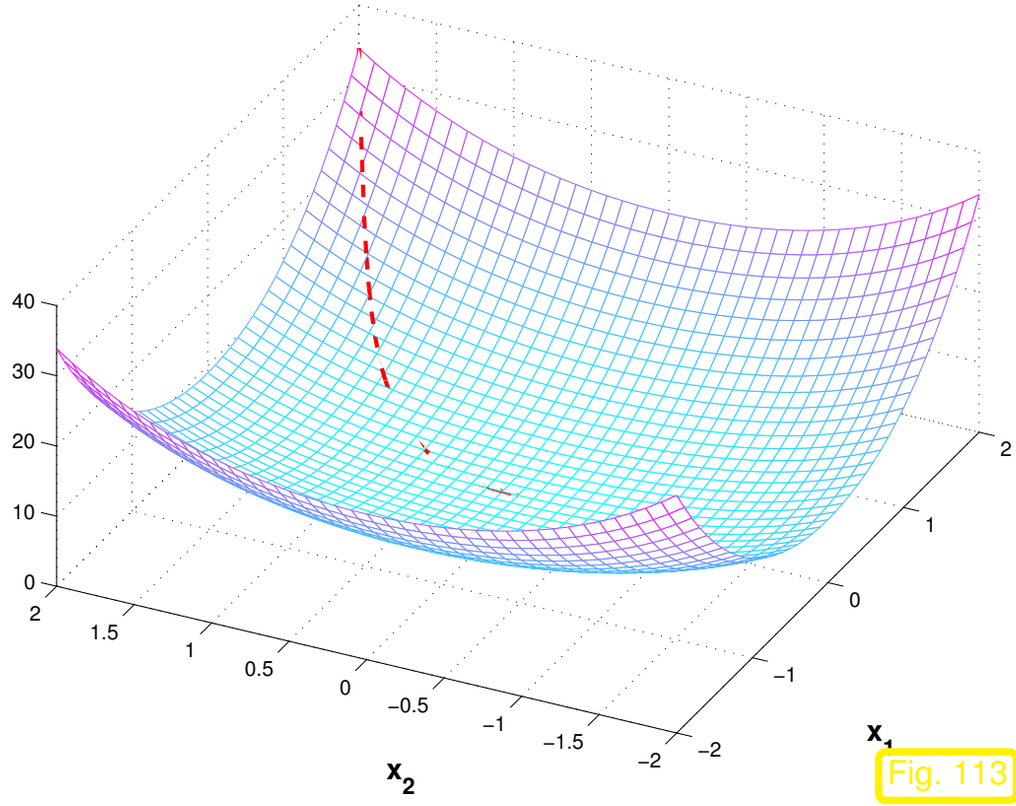


Fig. 113

Gradient-flow-IVP:

$$\begin{aligned} \dot{\mathbf{y}}(t) &= -\,\mathbf{grad}\,V(\mathbf{y}(t))\;, \\ \mathbf{y}(0) &= \mathbf{y}_0 \in \mathbb{R}^d\;. \end{aligned} \tag{3.3.4}$$

P. Grohs

rev 63606, February 17, 2014

The evolution to (3.3.4) is non-expansive rel. to the euclidean norm:

$$\chi(t) := \left\|\mathbf{\Phi}^t\mathbf{y} - \mathbf{\Phi}^t\widetilde{\mathbf{y}}\right\|_2^2 \quad\Rightarrow\quad \dot{\chi}(t) = -2\underbrace{(\mathbf{grad}\,V(\mathbf{\Phi}^t\mathbf{y}) - \mathbf{grad}\,V(\mathbf{\Phi}^t\widetilde{\mathbf{y}}))^T(\mathbf{\Phi}^t\mathbf{y} - \mathbf{\Phi}^t\widetilde{\mathbf{y}})}_{\geq 0} \leq 0\;.$$

➢ non-expansivity (→ Def. 3.3.1) with $\mathbf{M} = \mathbf{I}$.

Non-expansivity of an evolution: equivalent characterization

Definition 3.3.5 (dissipative vector field)**.**

$$\mathbf{f} : D \subset \mathbb{R}^d \mapsto \mathbb{R}^d \text{ dissipative} \quad :\Leftrightarrow \quad \mathbf{M}(\mathbf{f}(\mathbf{y}) - \mathbf{f}(\widetilde{\mathbf{y}})) \cdot (\mathbf{y} - \widetilde{\mathbf{y}}) \leq 0 \quad \forall \mathbf{y}, \widetilde{\mathbf{y}} \in D \,.$$

This is a generalization of the property "‘monotonically falling” of scalar valued functions.

Lemma 3.3.6 (Condition for non-expansivity of an evolution)**.**

Right hand side $\mathbf{f}$ *dissipative* $\Leftrightarrow$ *non-expansivity of the evolution to the autonomous ODE* $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$

Theorem 3.3.7 (Gauss-collocations-RK-OSM non-expansive)**.**
The discrete evolutions to Gauss-collocations-RK-OSM (→ Sect. 2.2.3) inherit the non-expansivity of the (exact) evolution.

Proof of Thm. 3.3.7. Consider Gauss-collocations-OSM with s nodes:

$\mathbf{y}_h(t), \widetilde{\mathbf{y}}_h(t) \in \mathcal{P}_s \hat{=}$ collocation-polynomial to initial values $\mathbf{y}_0$ resp. $\widetilde{\mathbf{y}}_0$, see Sect. 2.2.1

$$\blacktriangleright \qquad \mathbf{\Psi}^h \mathbf{y}_0 = \mathbf{y}_h(h) \quad , \quad \mathbf{\Psi}^h \widetilde{\mathbf{y}}_0 = \widetilde{\mathbf{y}}_h(h) \; .$$

$d(\tau) := \|\mathbf{y}_h(\tau h) - \widetilde{\mathbf{y}}_h(\tau h)\|_M^2$ is polynomial in τ of degree $\leq 2s$.

non-expansivity of $\mathbf{\Psi}^h$ is equivalent to

$$\left\|\mathbf{\Psi}^h \mathbf{y}_0 - \mathbf{\Psi}^h \widetilde{\mathbf{y}_0}\right\|_M^2 = d(1) = d(0) + \underbrace{\int_0^1 d'(\tau)\,\mathrm{d}\tau}_{\text{Ziel } \overset{!}{\leq} 0} = \|\mathbf{y}_0 - \widetilde{\mathbf{y}}_0\|_M^2 + \int_0^1 d'(\tau)\,\mathrm{d}\tau \; . \tag{3.3.8}$$

Gauss-quadrature (with s nodes) is exact for polynomials $\in \mathcal{P}_{2s-1}$

$$\blacktriangleright \qquad \int_0^1 d'(\tau)\,\mathrm{d}\tau = \sum_{j=1}^{s} b_j d'(c_j) \; . \tag{3.3.9}$$

Derivation from the chain-rule:

$$d'(\tau) = 2h\mathbf{M}(\mathbf{y}_h(\tau h) - \widetilde{\mathbf{y}}_h(\tau h)) \cdot (\dot{\mathbf{y}}_h(\tau h) - \dot{\widetilde{\mathbf{y}}}_h(\tau h)) \; . \tag{3.3.10}$$

From collocation-conditions (2.2.1):

$$\dot{\mathbf{y}}_h(c_j h) = \mathbf{f}(\mathbf{y}_h(c_j h)) \quad , \quad \dot{\widetilde{\mathbf{y}}}_h(c_j h) = \mathbf{f}(\widetilde{\mathbf{y}}_h(c_j h)) \; , \quad j = 1, \dots, s \; .$$

(3.3.10)

$$d'(c_j) = 2h\mathbf{M}(\mathbf{y}_h(c_j h) - \widetilde{\mathbf{y}}_h(c_j h)) \cdot (\mathbf{f}(\mathbf{y}_h(c_j h)) - \mathbf{f}(\widetilde{\mathbf{y}}_h(c_j h))) \leq 0 \; , \qquad (3.3.11)$$

since $\mathbf{f}$ dissipative $\Leftrightarrow$ non-expansivity of Φ^t, cf. Lemma 3.3.6.

(3.3.8), (3.3.9), (3.3.11) $\Rightarrow$ claim, since weights b_j of the Gauss-quadrature-formula is positive!. □

Lemma 3.3.12 (Discrete non-expansivity $\Rightarrow$ A-stability)**.**
non-expansivity ($\rightarrow$ Def. 3.3.1) inherits RK-OSM ($\rightarrow$ Def. 2.3.5) are A-stable ($\rightarrow$ Def. 3.2.16).

Proof. Scalar complex differential equation $\leftrightarrow$ real differential equation in $D = \mathbb{R}^2$: for arbitrary $\lambda = \alpha + i\beta \in \mathbb{C}$

$$\dot{y} = \lambda y \quad \overset{y=u+iv}{\Leftrightarrow} \quad \begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \underbrace{\begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}}_{=:\mathbf{A}} \begin{pmatrix} u \\ v \end{pmatrix} \quad \leftrightarrow \quad \dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) \text{ with } \mathbf{y} := \begin{pmatrix} u \\ v \end{pmatrix} \; .$$

$$\operatorname{Re}\lambda < 0 \quad \Rightarrow \quad \alpha < 0 \quad \Rightarrow \quad \mathbf{x}^T\mathbf{A}\mathbf{x} = \alpha\,\|\mathbf{x}\|_2 \leq 0 \quad \forall \mathbf{x} \in \mathbb{R}^2$$

$\Rightarrow$ Right hand side $\mathbf{f}(\mathbf{y}) = \mathbf{A}\mathbf{y}$ is dissipative ($\rightarrow$ Def. 3.3.5)

$\Rightarrow$ Evolution non-expansive, see Lemma 3.3.6.

("Inheritance") ▶ discrete evolution $\mathbf{\Psi}^h$ non-expansive

$$\Rightarrow \quad \left\|\mathbf{\Psi}^h\mathbf{y}\right\|_2 = |S(h\lambda)||y| \leq \|\mathbf{y}\|_2 = |y| \quad \Rightarrow \quad |S(z)| \leq 1 \quad \forall z \in \overline{\mathbb{C}^-}\,.$$

$$\overset{*}{\Rightarrow} \quad |S(z)| < 1 \quad \forall z \in \mathbb{C}^-\,.$$

$$\Rightarrow \quad \mathbb{C}^- \subset \mathcal{S}_{\mathbf{\Psi}} \quad \text{(stability domain} \rightarrow \text{Def. 3.1.4)}\,. \qquad \square$$

$*$: $S(z)$ is holomorphic on $\mathbb{C}^-$, such that (maximum principle!) $|S(z)|$ can attain its maximum on $\overline{\mathbb{C}}^-$ only on the boundary $\partial\mathbb{C} = i\mathbb{R}$.

P. Grohs

rev 63606, February 17, 2014

▶ Alle Gaus-collocations-OSM are A-stable

Remark 3.3.13 (Solvability of the incremental equation for Gauss-collocations-OSM).

The incremental equation of a Gauss-collocations-RK-OSM for a non-expansive autonomous ODE are uniquely solvable for each $h > 0$ $\rightarrow$[23, Sect. IV.14].

Stability domain of Gauss-collocations-one-step-method:

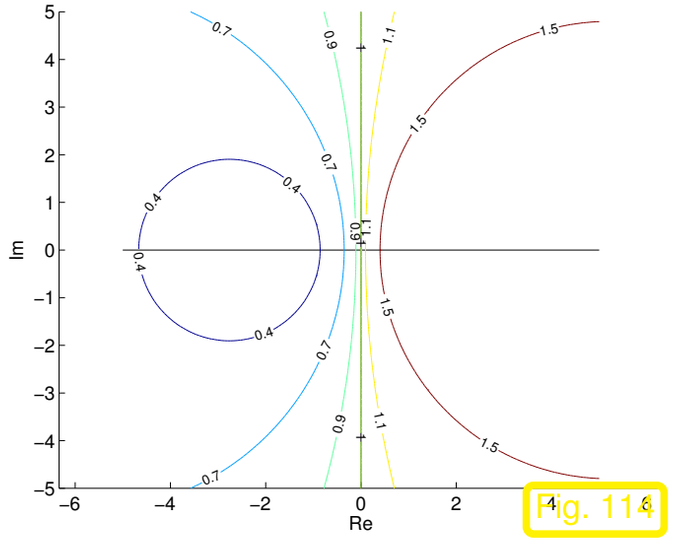


Fig. 114

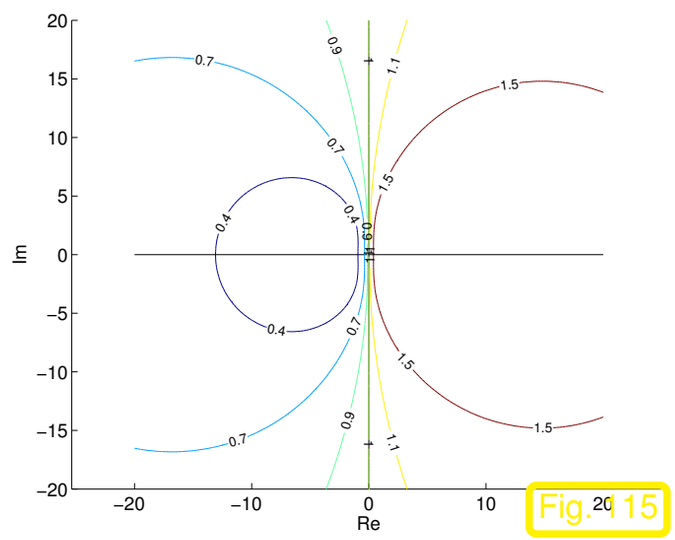


Fig. 115

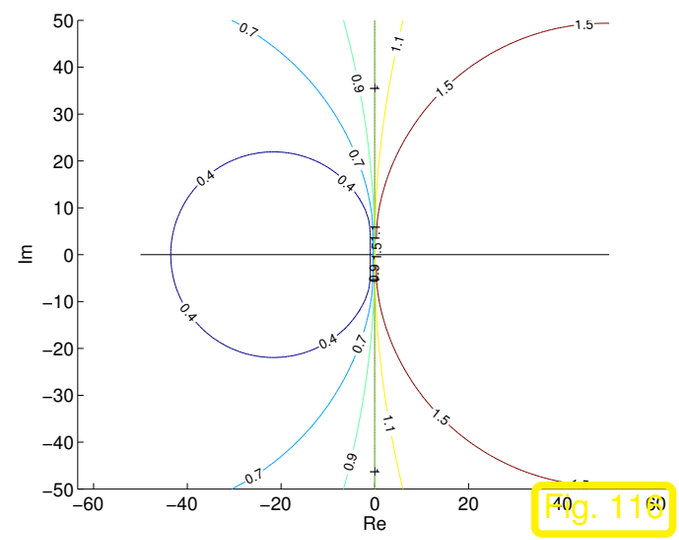


Fig. 116

Implicit midpoint-rule $\quad$ $s = 2$ (Order 4) $\quad$ $s = 4$ (Order 8)

Level-lines of $|S(z)|$ for Gauss-collocations-one-step-method

Conjecture (Proof later): $\mathcal{S}_{\Psi} = \mathbb{C}^{-}$

Example 3.3.14 (Gauss-collocation-method for logistic differential equation). $\rightarrow$ Bsp. 3.0.1, 1.4.21

logistic differential equation $\dot{y} = f(y)$, $f(y) = \lambda y(1-y) \to$ (2.2.57), $\lambda = 50$, initial value $y_0 = 10^{-4}$, time interval $[0,1]$.

collocations-one-step-method ($\to$ Section 2.2) on equidistant mesh, $h = \frac{1}{20}$.

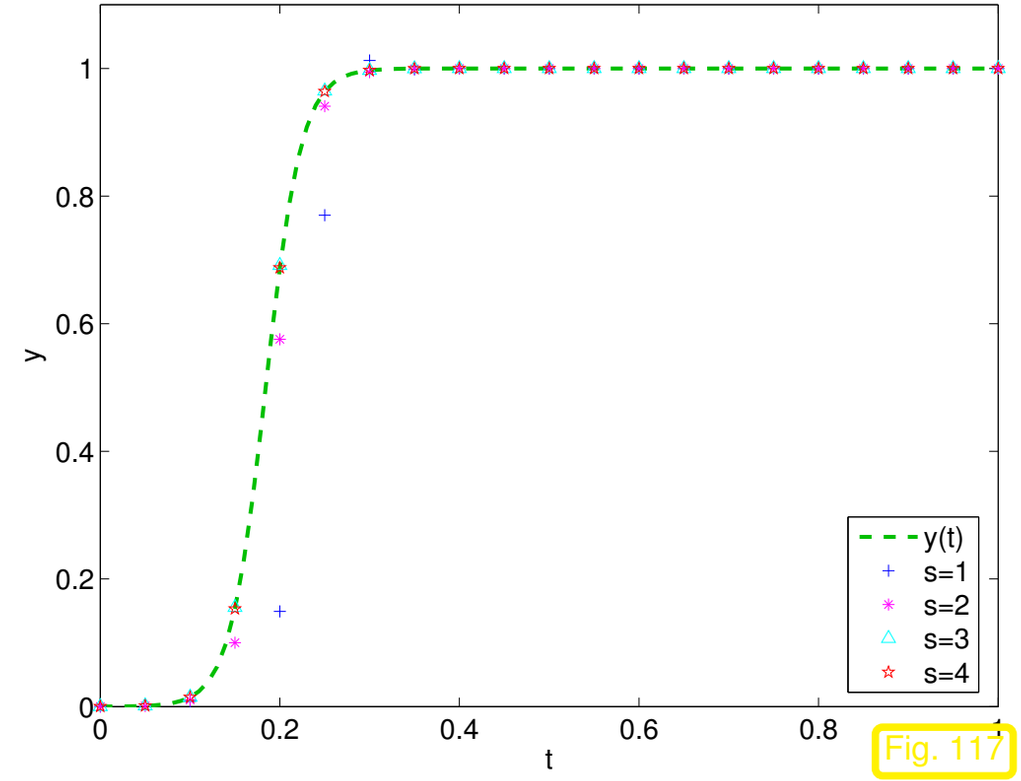


Fig. 117

◇

P. Grohs

rev 63606, February 17, 2014

Remark 3.3.15 (A-stability $\not\Rightarrow$ discrete non-expansivity).

Counter example: implicit trapezoidal rule, one-step-method for $\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y})$ defined by

$$\mathbf{y}_1 = \mathbf{y}_0 + \tfrac{1}{2}h(\mathbf{f}(t,\mathbf{y}_0) + \mathbf{f}(t+h,\mathbf{y}_1)) \quad \leftrightarrow \quad \text{Butcher-Schema} \quad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}$$

applied on scalar autonomous ODE $\dot{y} = \begin{cases} -y^3 & \text{für } y < 0 \,, \\ -y^2 & \text{für } y \geq 0 \,. \end{cases}$ $\rightarrow$ exercise

Remark 3.3.16 (B-stability).

One-step-method, which inherits the non-expansivity of the evolution of an ODE are also called B-stable [23, Sect. IV.12].

△

An algebraic criteria for B-stability:

Definition 3.3.17 (algebraic stability)**.**

A Runge-Kutta-one-step-method ($\rightarrow$ Def. 2.3.5) with Butcher-Schema $\begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^T \end{array}$*, see* (2.3.6)*, is algebraic stable if*

(i) $b_i \geq 0$, $i = 1, \dots, s$,

(ii) and the matrix $\mathbf{M} := \operatorname{diag}(b_1, \dots, b_s)\mathfrak{A} + \mathfrak{A}^T \operatorname{diag}(b_1, \dots, b_s) - \mathbf{b}\mathbf{b}^T$ *is positive semi-definite.*

Theorem 3.3.18 (criteria for B-stability)**.**

$$\textit{algebraic stability} \quad \Rightarrow \quad \textit{B-stability}$$

3.4 Uniform stability

Example 3.4.1 (Gauss-collocations-OSM with strongly attractive fixpoints). $\rightarrow$ Ex. 3.5.2

ODE with strongly attractive fixpoint $y = 1$:

$$\dot{y} = \lambda y^2(1-y) \ ,$$
$$\lambda = 500 \ , \quad y(0) = \frac{1}{100} \ .$$

qualitative behavior of Gauss-collocations-OSM ($\rightarrow$ Section 2.2) on equidistant mesh $\triangleright$

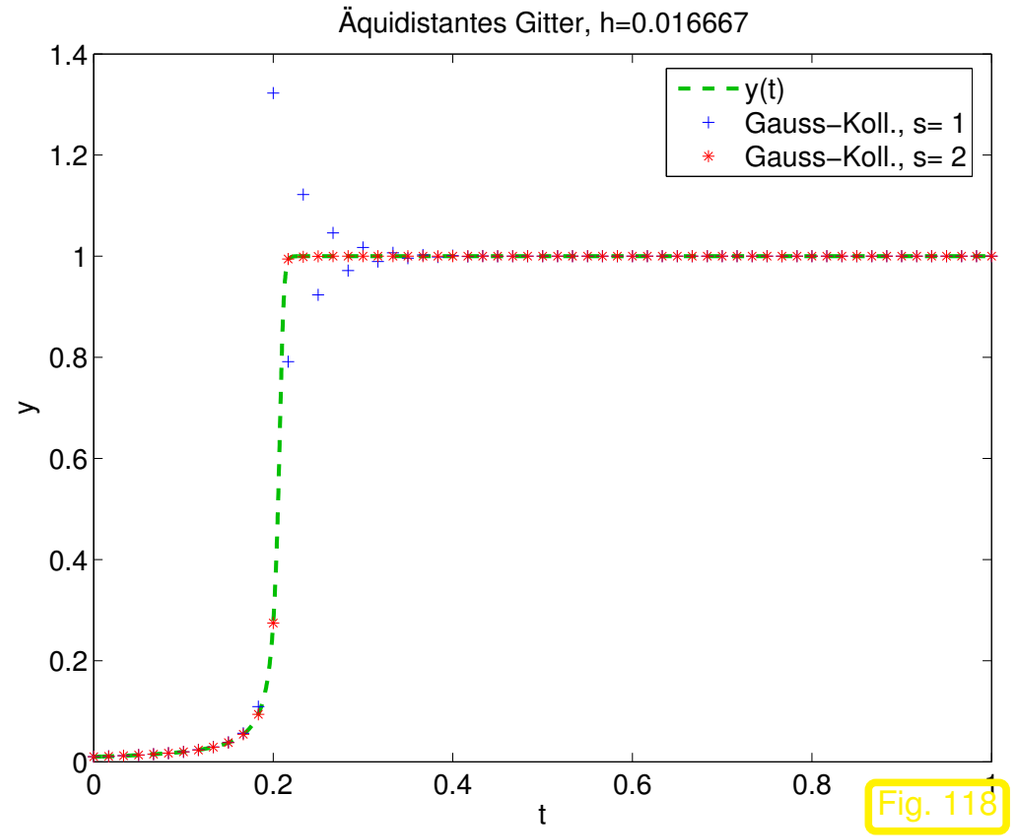


Fig. 118

P. Grohs

rev 63606, February 17, 2014

➢ Wrong oscillations with Gauss-collocations-OSM of low order

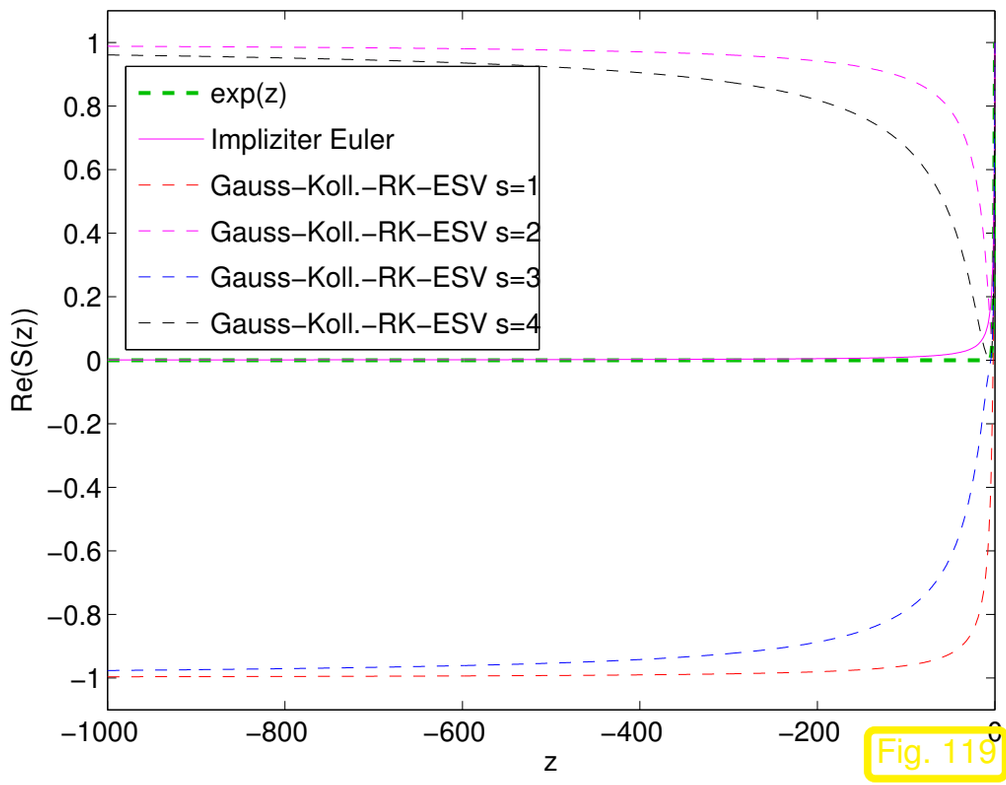
Explanation: For Gauss-collocations-OSM we have $S(z) \approx \pm 1$ for $|z| \rightarrow \infty$ such that the fixpoint of the discrete evolution stays attracting but the discrete solution (in contrast to the continuous) only slowly (and convergences oscillatori for odd s to it.)

Further demonstration $\rightarrow$ Ex. 3.4.2

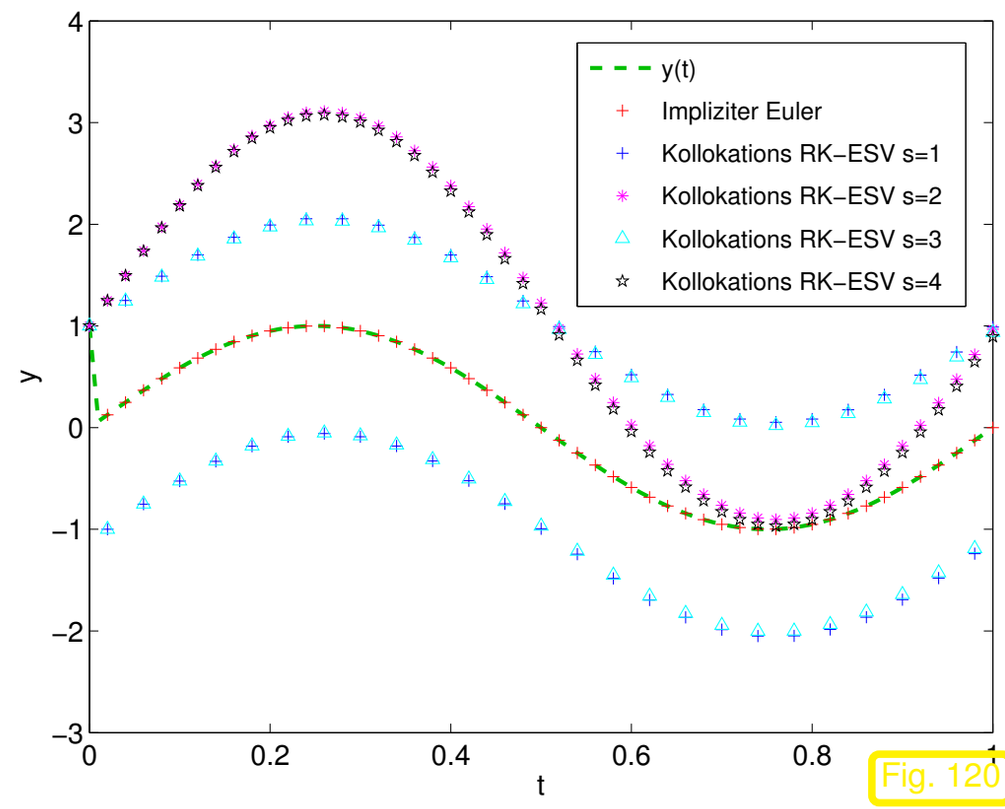
Example 3.4.2 (Implicit RK-OSM with fast transients). → Ex. 3.5.5

IVP: $\dot{y} = -\lambda y + \beta \sin(2\pi t)$, $\lambda = 10^6$, $\beta = 10^6$, $y(0) = 1$.

RK-OSM, equidistant mesh on $[0,1]$, $h = \frac{1}{40}$:



Stability functions for $\mathrm{Re}\, z \ll 1$



Discrete evolutions (time elapsed)

➢ Insufficient absorption of the initial perturbation with collocations-RK-OSM **!**
(Oscillations for odd s ➞ cf. stability functions, $\lim\limits_{\mathrm{Re}\, z \to -\infty} S(z) = (-1)^s$)

➢ Implicit euler-method (1.4.13): *immediate* relaxation of the discrete solution **!**

Clear, because $\lim\limits_{\mathrm{Re}\, z \to -\infty} S(z) = \lim\limits_{\mathrm{Re}\, z \to -\infty} \frac{1}{1-z} = 0$ for implicit euler-method.

◇

P. Grohs

rev 63606, February 17, 2014

Sect. 3.1: $\boxed{\text{stability function } S(z) \quad \longleftrightarrow \quad \exp(z)}$

Utopia (for RK-OSM): $S(-\infty) = 0 \quad , \quad S(\infty) = \infty$

(can not be fullfilled by a rational function **!**)

Modest wish (for strongly attractive fixpoints, fast relaxations): $S(-\infty) = 0$

Definition 3.4.3 (L-stability)**.**

$$\textit{OSM L-stable} \quad :\Leftrightarrow \quad \{z \in \mathbb{C}\colon \operatorname{Re} z < 0\} \subset \mathcal{S}_\Psi \quad \& \quad \lim_{\operatorname{Re} z \to -\infty} |S(z)| = 0$$

Short: L-stable $:\Leftrightarrow$ A-stable **&** "$S(-\infty) = 0$"

How do we find L-stable RK-OSM **?** Existence is required

$$\text{Thm. 3.1.6} \quad \Rightarrow \quad S(-\infty) = 1 - \mathbf{b}^T \mathfrak{A}^{-1} \mathbf{1} \ . \tag{3.4.4}$$

$$\boxed{\text{If } \mathbf{b}^T = \mathbf{a}_{j,\cdot}^T \text{ (Row of } \mathfrak{A}) \quad \wedge \quad \mathfrak{A} \text{ regular} \quad \Rightarrow \quad S(-\infty) = 0} \ . \tag{3.4.5}$$

characterized stiffly accurate RK-OSM [12, Lemma 6.32]

P. Grohs

rev 63606, February 17, 2014

Remark 3.4.6 (Invertibility of the coefficient matrix of RK-OSM).

For each s-level collocation method ($\rightarrow$ Sect. 2.2.1) with $c_s > 0$ (i.e. for each collocation method with exception of the explicit euler-method (1.4.2)) the coefficient matrix (Butcher-Matrix) $\mathfrak{A}$ is non-singular

Proof. Let $\mathbf{x} \in \mathbb{R}^s$ with $\mathfrak{A}\mathbf{x} = 0$

$$\overset{(2.2.3)}{\Rightarrow} \quad \sum_{j=1}^{s} a_{ij} x_j = \sum_{j=1}^{s} \int_0^{c_i} x_j L_j(\tau)\,\mathrm{d}\tau = 0\,, \quad i = 1, \dots, s\,,$$

with the Lagrange-polynomial $L_i \in \mathcal{P}_{s-1}$ from (2.2.2).

$$\blacktriangleright \quad q := \sum_{j=1}^{s} x_j L_j \quad \text{satisfied:} \quad \int_{c_{i-1}}^{c_i} q(\tau)\,\mathrm{d}\tau = 0\,, \quad i = 1, \dots, s\ (c_0 := 0)\,.$$

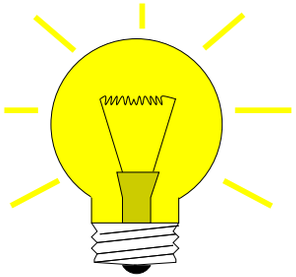
$\Rightarrow \quad q \in \mathcal{P}_{s-1}$ has s roots in $[0, c_s] \quad \Rightarrow \quad q = 0 \quad \Rightarrow \quad \mathbf{x} = 0$. $\square$ $\triangle$

P. Grohs

rev 63606, February 17, 2014

Butcher-Scheme (2.3.6) for consistent ($\rightarrow$ Lemma 2.3.23), L-stable RK-OSM, see Def. 3.4.3

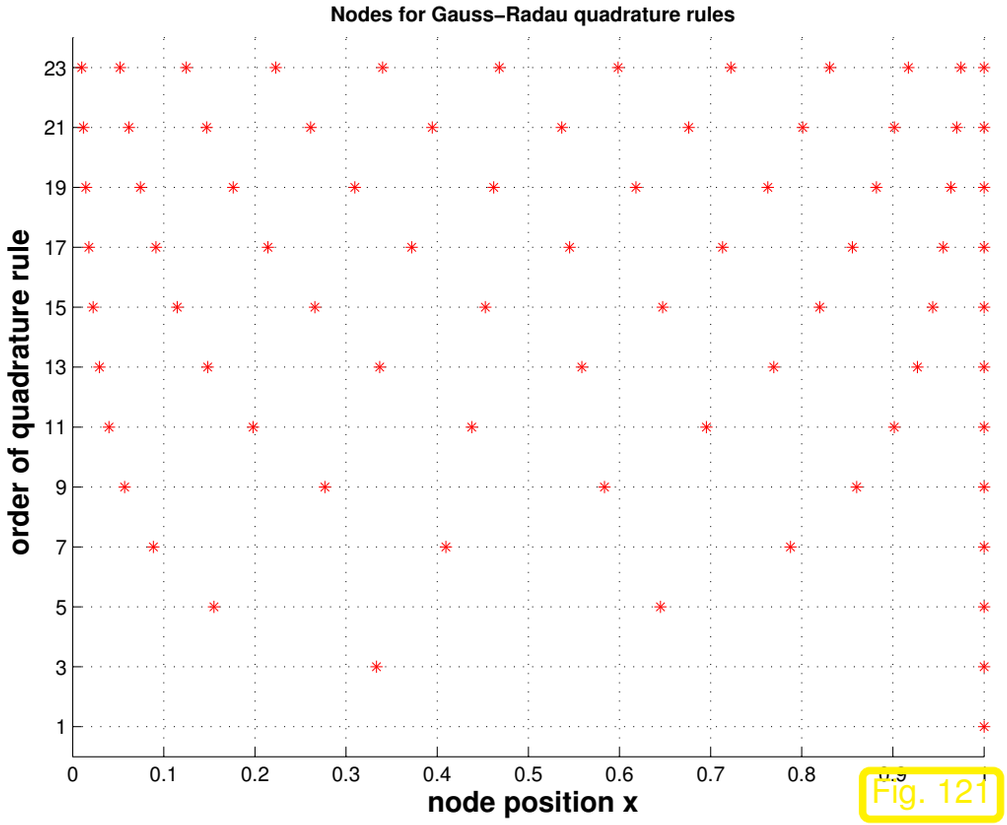
$$\triangleright \quad \begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^T \end{array} := \begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_{s-1} & a_{s-1,1} & \cdots & a_{s-1,s} \\ 1 & b_1 & \cdots & b_s \\ \hline & b_1 & \cdots & b_s \end{array} .$$



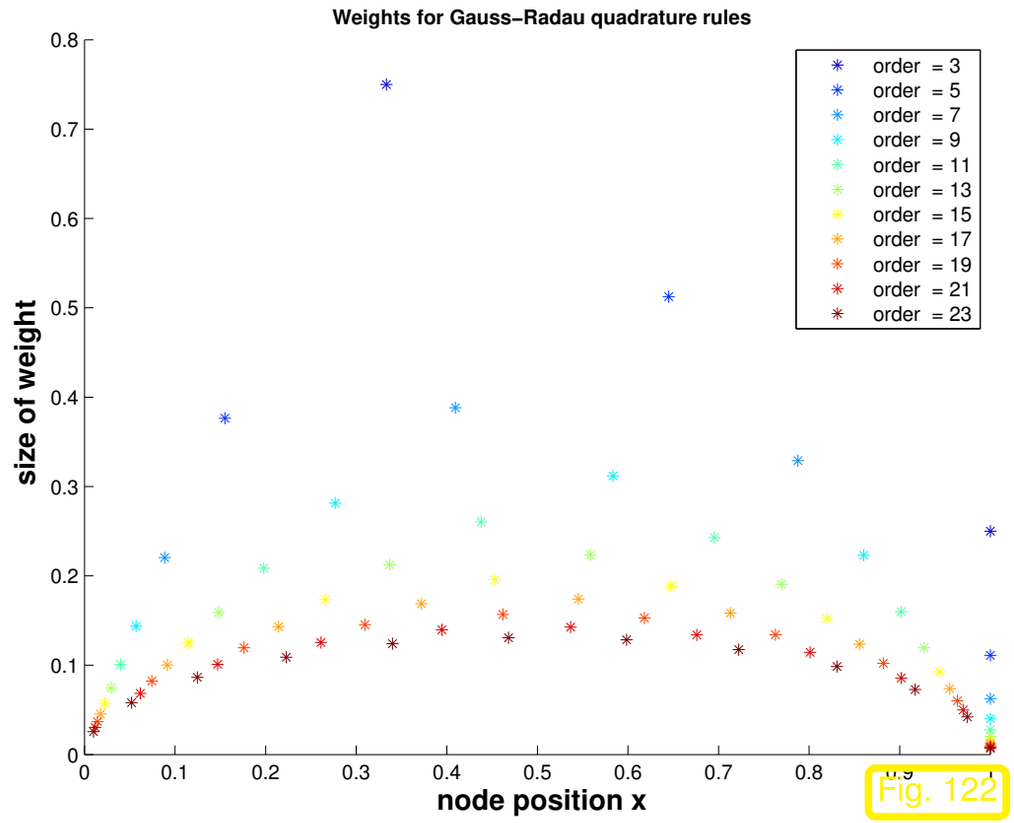
Idea: Choose $c_s = 1$ in the collocations-RK-OSM (2.2.3)

Choose $c_1, \ldots, c_{s-1}$ as nodes of a quadrature formula of maximal order.

($\rightarrow$ Gauss-Radau-quadrature, order $2s - 1$)



Nodes: Radau-quadrature-formula



Weights: Radau-quadrature-formula

P. Grohs

rev 63606, February 17, 2014

Implicit s-level L-stable Radau-OSM, convergence order $2s-1$

($\rightarrow$ Thm. 2.2.50, [12, Sect. 6.3.2])

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

implicit Euler-OSM

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}$$

Radau-OSM, order 3

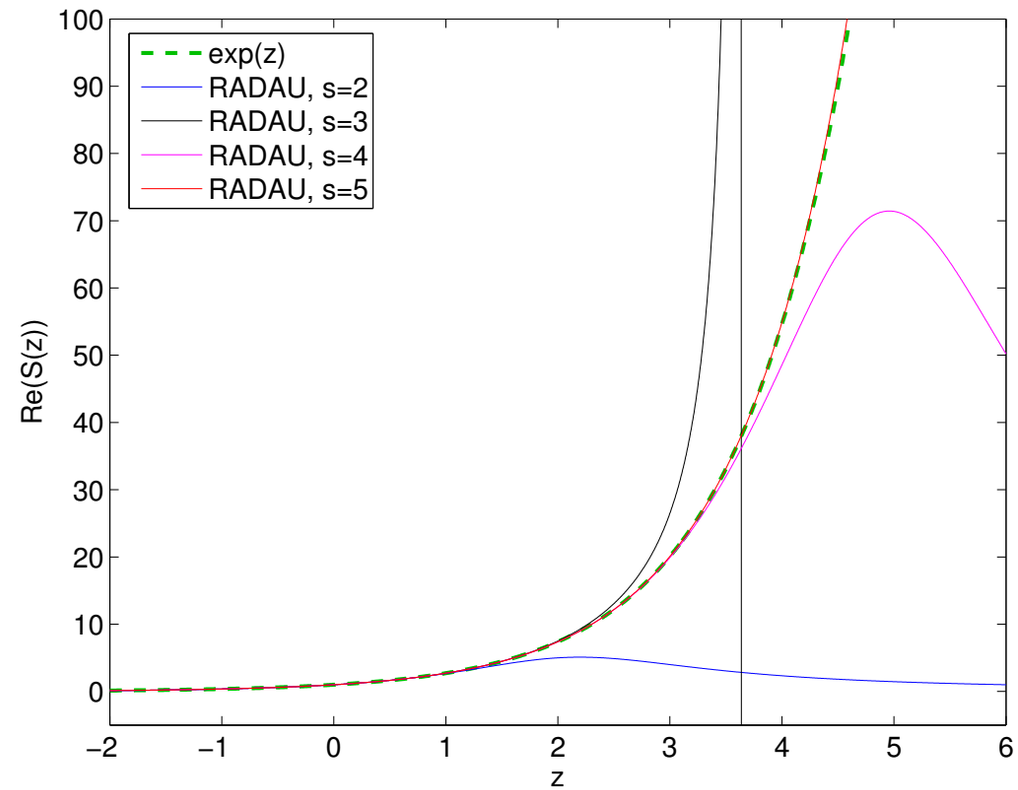
$$\begin{array}{c|ccc} \frac{4-\sqrt{6}}{10} & \frac{88-7\sqrt{6}}{360} & \frac{296-169\sqrt{6}}{1800} & \frac{-2+3\sqrt{6}}{225} \\ \frac{4+\sqrt{6}}{10} & \frac{296+169\sqrt{6}}{1800} & \frac{88+7\sqrt{6}}{360} & \frac{-2-3\sqrt{6}}{225} \\ 1 & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \\ \hline & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \end{array}$$

Radau-OSM, order 5

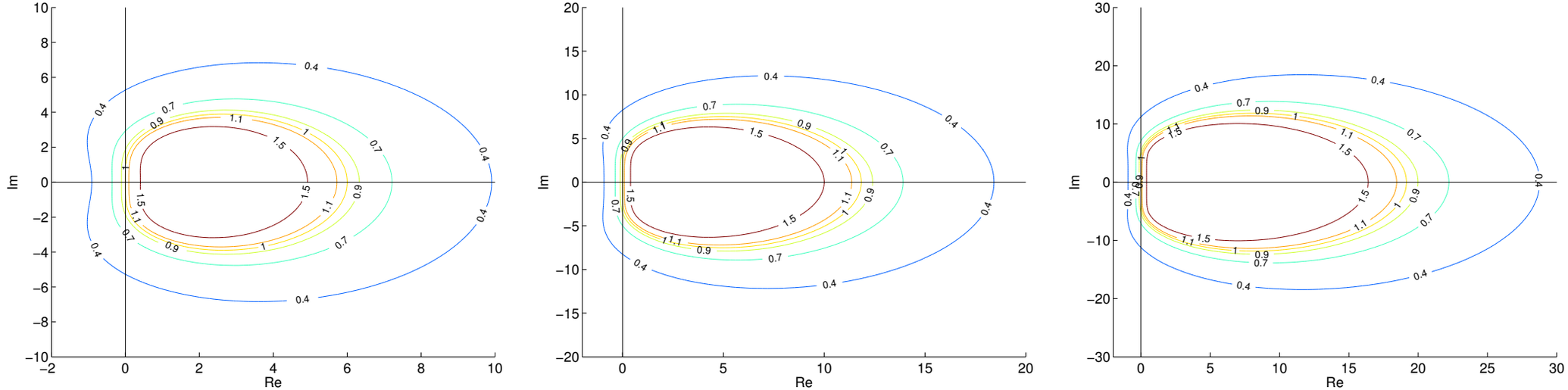
stability function s-level Radau-collocations-RK-OSMs:

$$S(z) = \frac{P(z)}{Q(z)}\,, \quad P \in \mathcal{P}_{s-1}, Q \in \mathcal{P}_s\,.$$

Attention: Also "‘$S(\infty) = 0$”



Levels of the stability functions of s-level Radau-collocations-RK-OSMs:



$s = 2$ $s = 3$ $s = 4$

Example 3.4.7 (Radau-OSM with strong attractive fixpoints). $\rightarrow$ Ex. 3.4.1

IVP for ODE with strong attractive fixpoint $y = 1$, Ex. 3.4.1

$$\dot{y} = \lambda y^2(1-y) \;,$$
$$\lambda = 500 \;, \quad y(0) = \frac{1}{100} \;.$$

Qualitative behavior of Radau-OSM on equidistant mesh ▷

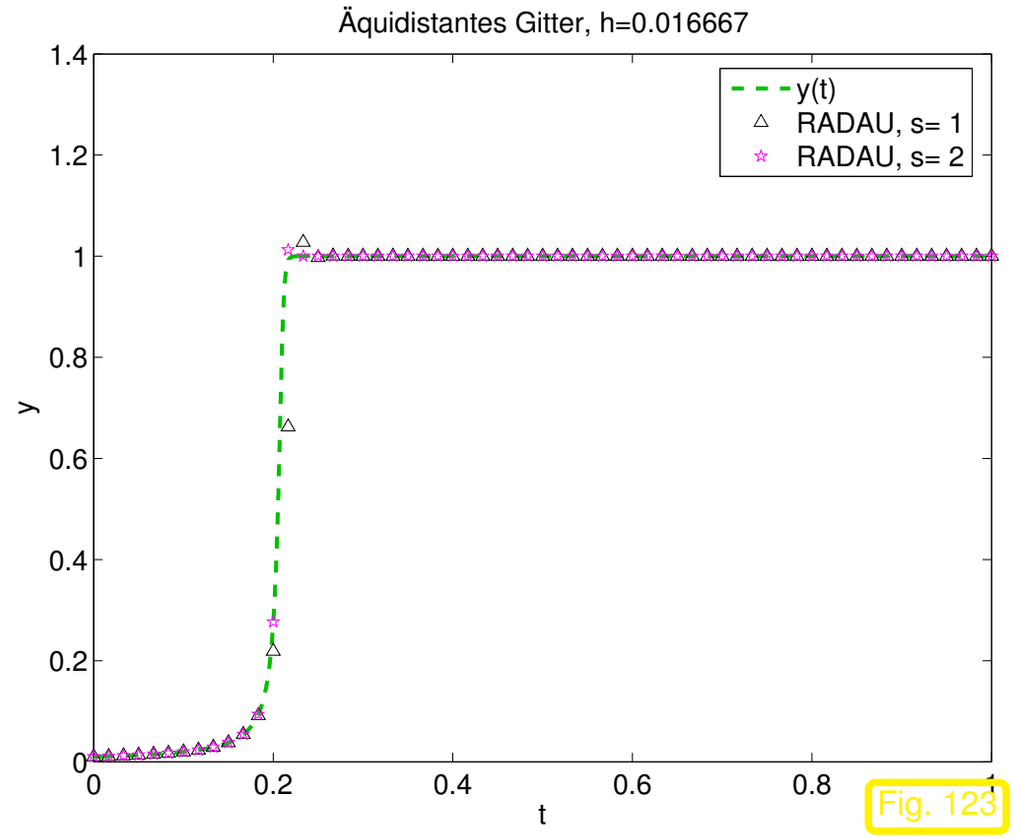


Fig. 123

P. Grohs

rev 63606, February 17, 2014

➢ discrete evolution converges fast to the stable state

◇

As for Gauss-collocations-RK-OSM, see Thm. 3.3.7:

Theorem 3.4.8 (Radau-OSM non-expansive)**.**
The discrete evolutions to Radau-OSM inherits the non-expansivity of the (exact) evolution.

Proof. Extension of the proof of Thm. 3.3.7. With the notations and error representation formula there for Gauss-Radau-quadrature formula

$$\int_0^1 d'(\tau)\,\mathrm{d}\tau = \sum_{j=1}^{s} b_j d'(c_j) - R \quad \text{mit} \quad R = c(s) d^{(2s)}(\xi)\ , \quad 0 \le \xi \le 1\ ,$$

where $c(s) > 0$. Formula (3.3.10) and (3.3.11) stay valid such that the claim of Thm. 3.4.8 is shown as soon as $R \ge 0$ is guaranteed:

$$d(\tau) = \sum_{j=0}^{2s} \delta_j \tau^j \quad \Rightarrow \quad d^{(2s)}(\tau) = (2s)!\delta_{2s}\ ,$$

$$d(\tau) \ge 0 \quad \Rightarrow \quad \lim_{|\tau|\to\infty} d(\tau) \ge 0 \quad \Rightarrow \quad \delta_{2s} \ge 0\ .$$

Note: Also the weights of Gauss-Radau-quadrature formula are positive, see Fig. 122. □

Radau-OSM are A-stable (→ Def. 3.2.16)

3.5 Stiffness

The following is not a definition but a term motivated by observations of user of numerical integrators. It is not possible to give a rigorous mathematical definition of this term. Nevertheless it is central for the choice of adequate numerical integrators and after all attenders of the lecture should get a “feeling” if an initial value problem is “stiff”. This will be trained in this section with the help of examples.

Aus [30, Sect. 1]:

> The usual definition of stiffness applies which states that a differential equation is stiff whenever the implicit Euler method works (tremendously) better than the explicit Euler method.

Notion 3.5.1 (Stiff initial value problem)**.**
An IVP is called stiff if for explicit RK-OSM (→ Def. 2.1.5) stability requires a significantly smaller stepwidth than the precision conditions.

Example 3.5.2 (Adaptive explicit RK-OSM for stiff problem). → Sect. 2.6

$$\dot{y}(t) = \lambda y^2(1-y)\ , \quad \lambda = 500 \quad , \quad y(0) = \frac{1}{100}\ .$$

MATLAB-CODE : Adaptive OSM for stiff problem

```
fun = @(t,x) 500*x^2*(1-x); tspan = [0 1]; y0 = 0.01;
options = odeset('reltol',0.1,'abstol',0.001,'stats','on');
[t,y] = ode45(fun,tspan,y0,options);
plot(t,y,'r+');
```

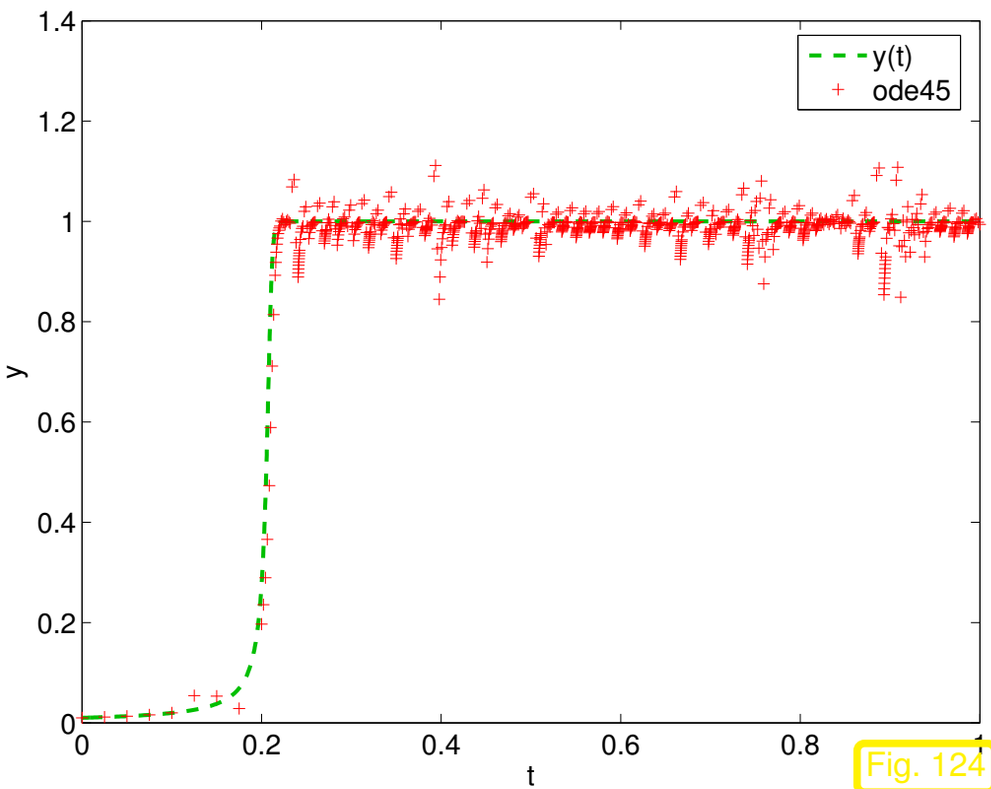


Fig. 124

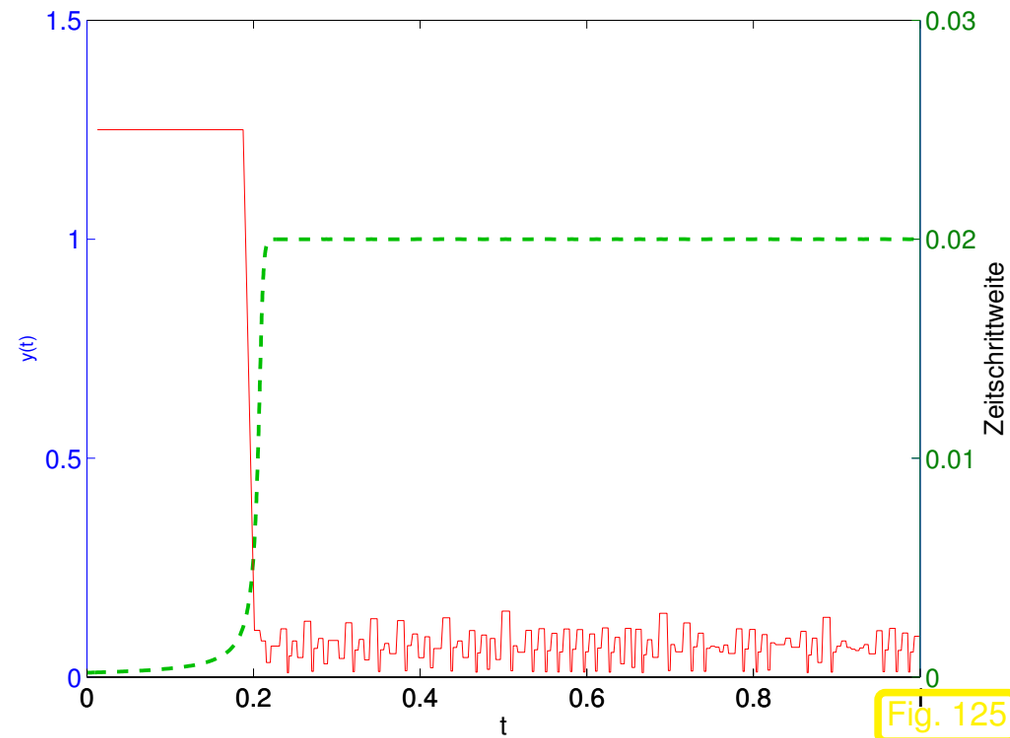


Fig. 125

➢ Control of stepwidth realized stepwidth condition ! → Ex. 2.6.10

(186 successful steps, 55 failed attempts, 1447 function evaluations)

$\boxed{y = 1 \text{ strong attractive fixpoint}}$ ➢ Extreme stepwidth condition for explicit integrator `ode45`

Note: the control of stepwidth recognized stability problems and reduces the stepwidth accordingly !
→ Ex. 2.6.10

◇

Which initial value problems are stiff **?**

☞ ODE-models for systems with fast relaxing components
(with highly unequal time constants)

Example 3.5.3 (Stiff problems in chemical reaction kinetics). → Sect. 1.2.2

Reaktion: $$\mathrm{A} + \mathrm{B} \underset{k_1}{\overset{k_2}{\rightleftharpoons}} \mathrm{C} \quad , \quad \mathrm{A} + \mathrm{C} \underset{k_3}{\overset{k_4}{\rightleftharpoons}} \mathrm{D} \tag{3.5.4}$$

Highly unequal reaction velocities: $\boxed{k_1, k_2 \gg k_3, k_4}$

If $c_A(0) > c_B(0)$ ➢ 2. reaction controlles longtime dynamics

numerical experiment, MATLAB, $t_0 = 0$, $T = 1$, $k_1 = 10^4$, $k_2 = 10^3$, $k_3 = 10$, $k_4 = 1$

MATLAB-CODE : Explicit integration of stiff chemical reaction equation

```
fun = @(t,y) ([-k1*y(1)*y(2) + k2*y(3) - k3*y(1)*y(3) + k4*y(4);
              -k1*y(1)*y(2) + k2*y(3);
              k1*y(1)*y(2) - k2*y(3) - k3*y(1)*y(3) + k4*y(4);
              k3*y(1)*y(3) - k4*y(4)]);
tspan = [0 1];
y0 = [1;1;10;0];
options = odeset('reltol',0.1,'abstol',0.001,'stats','on');
[t,y] = ode45(fun,[0 1],y0,options);
```

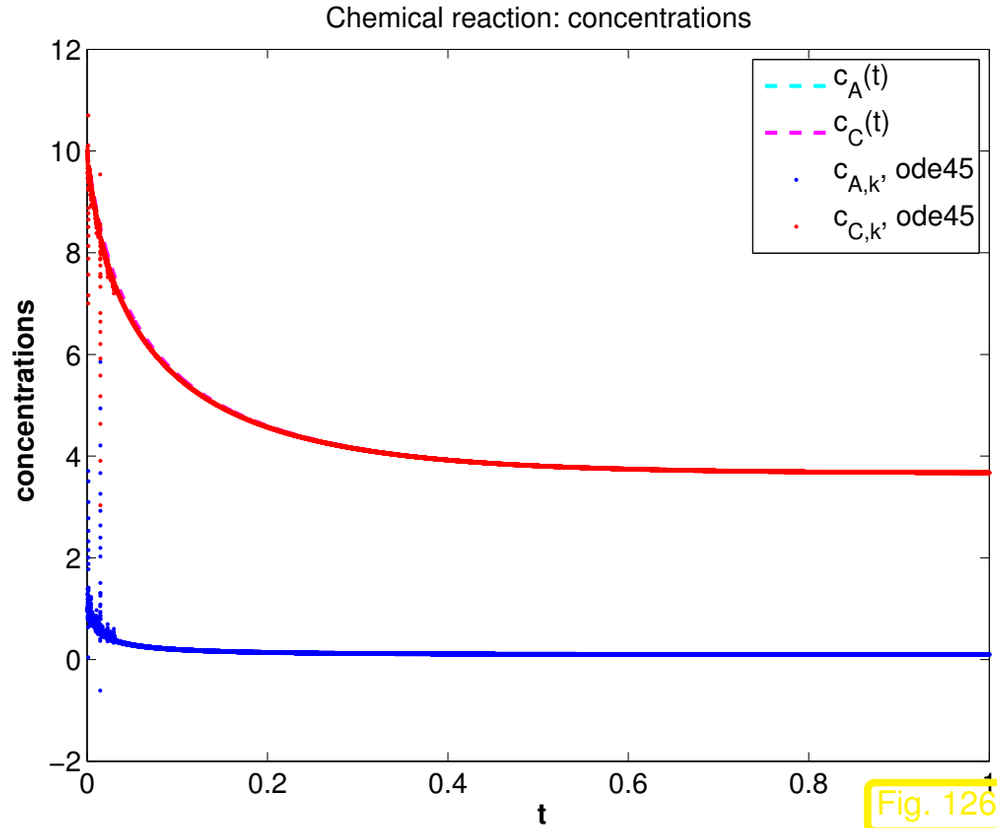


Fig. 126

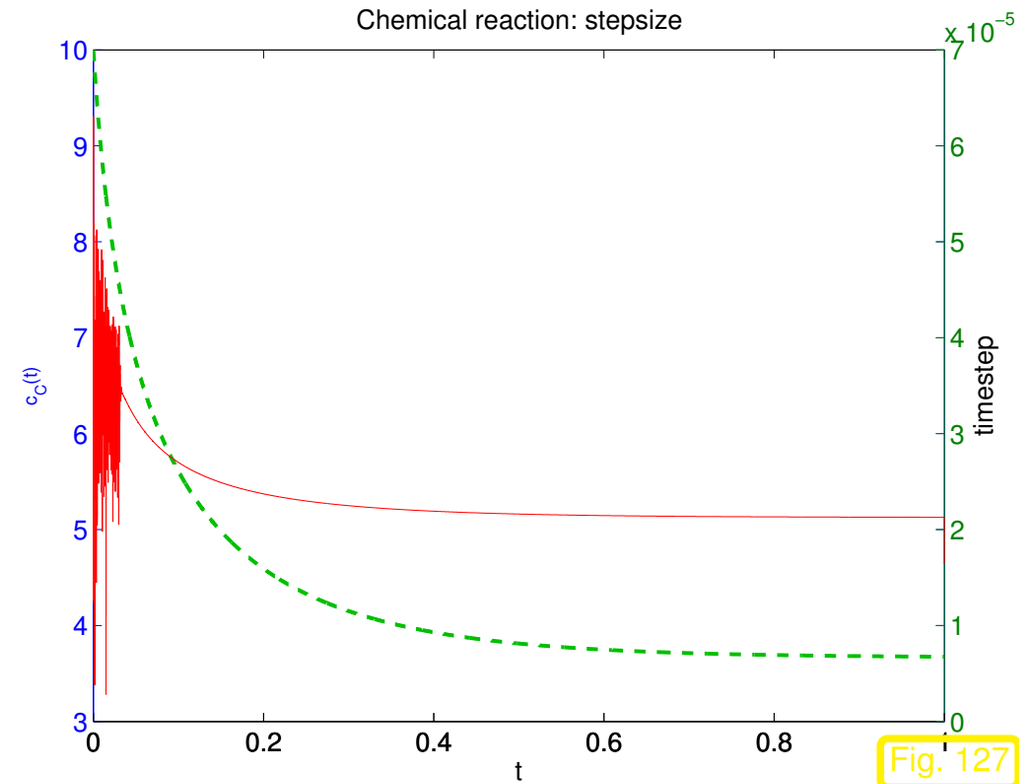


Fig. 127

◇

Example 3.5.5 (Stiff circuit equations in the time domain).

P. Grohs

rev 63606, February 17, 2014

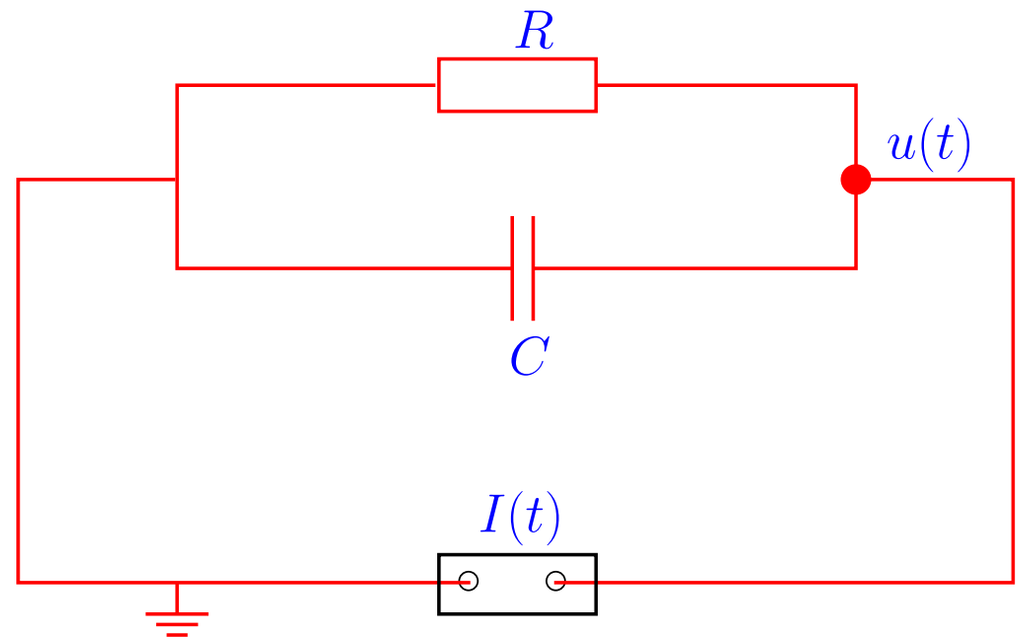
Circuit analysis in the time domain:

Building panel equation ($i \hat{=}$ electricity, $u \hat{=}$ voltage):

- Resistance: $i(t) = R^{-1}u(t)$
- Capacitor: $i(t) = C\dot{u}(t)$

Differential equation from node analysis:

$$C\dot{u}(t) = -R^{-1}u(t) + I(t) \ .$$



Concrete: $C = 1pF$, $R = 1k\Omega$, $I(t) = \sin(2\pi\, 1\mathrm{Hz}\, t)\mathrm{mA}$, $u(0) = 0V$

► Scaled (dimensionsless) differential equation: $\dot{u}(t) = -10^9\, u(t) + 10^9 \sin(2\pi t) \quad \Rightarrow \quad u(t) \approx$

$\hat{=}$ ODE from Ex. 3.4.2

In the case of the non-autonomous differential equation $\dot{y} = -\lambda y + g(t)$, $\lambda \gg 1$, we are confronted with a "timely varying " highly attractive fixpoint $y^*(t) = \lambda^{-1}g(t)$. Also this leads to stiffness according to concept 3.5.1.

Stiff IVP ➤ (appropriate) implicit RK-OSM ($\rightarrow$ Def. 2.1.5) are to be used !

d.h. L-stable ($\rightarrow$ Def. 3.4.3)

Example 3.5.6 (Attractive limit cycle). cf. Ex. 1.2.15

autonomous differential equation $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$

$$\mathbf{f}(\mathbf{y}) := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \mathbf{y} + \lambda(1 - \|\mathbf{y}\|^2)\,\mathbf{y}\ ,$$

on state space $D = \mathbb{R}^2 \setminus \{0\}$.

trajectories of solutions ($\lambda = 10$) ▷

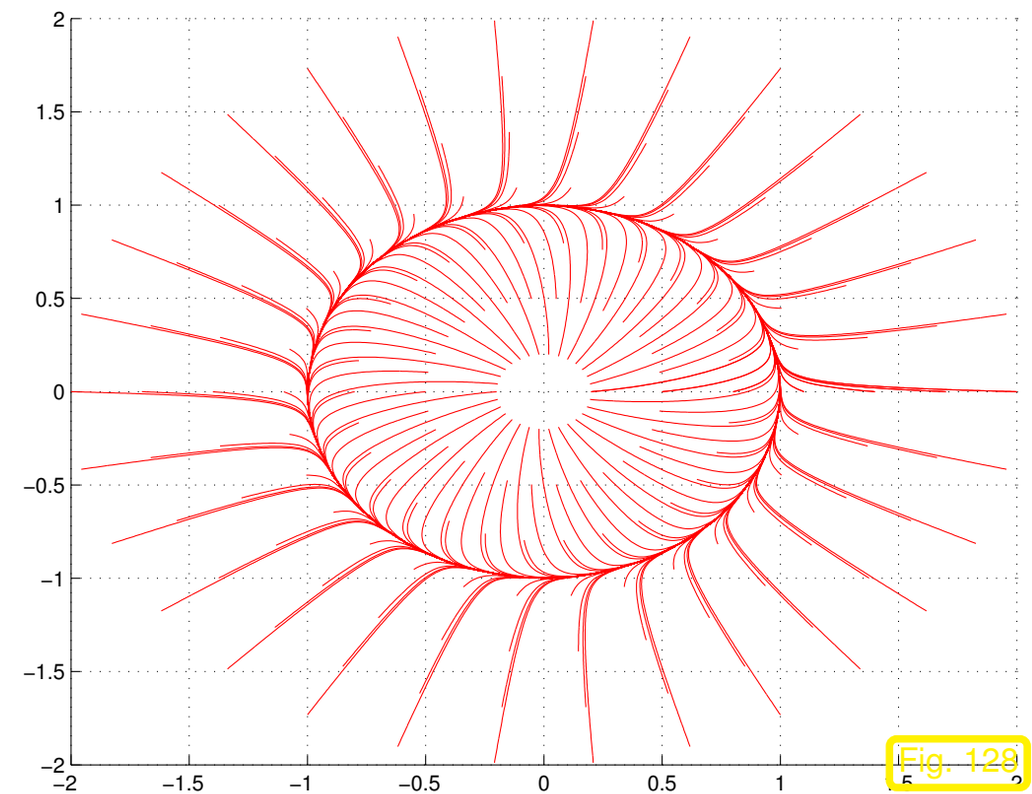


Fig. 129

☞ Falls $\|\mathbf{y}_0\| = 1 \Rightarrow \|\mathbf{y}(t)\| = 1 \quad \forall t$

☞ "Limit cycles on unit circles": $\|\mathbf{y}(t)\| \to 1$ for $t \to \infty$.

In this example there is no asymptotic stable fixpoint but a asymptotic stable invariant manifold, also a proper subset $M \subset D$ of the state space, for which $\Phi^t M \subset M$ holds for all admissible t ("fixed set") and

$$\exists \text{Neighbourhood } U \text{ of } M: : \quad \mathbf{y}(0) \in U \quad \Rightarrow \quad \lim_{t\to\infty} \operatorname{dist}(\mathbf{y}(t), M) = 0 \ .$$

MATLAB-CODE Integration of evolution with limit cycle

```
fun = @(t,y) ([-y(2);y(1)] + lambda*(1-y(1)^2-y(2)^2)*y);
tspan = [0,2*pi]; y0 = [1,0];
opts = odeset('stats','on','reltol',1E-4,'abstol',1E-4);
[t45,y45] = ode45(fun,tspan,y0,opts);
[t23,y23] = ode23s(fun,tspan,y0,opts);
```

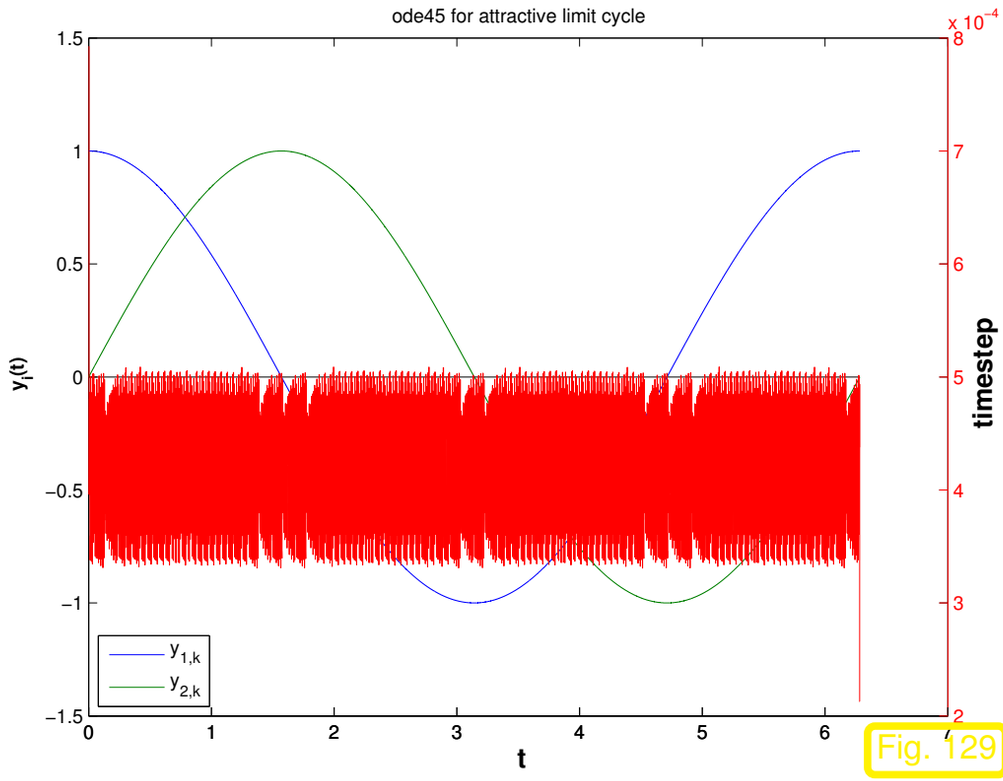


Fig. 129

`ode45`: 3794 steps ($\lambda = 1000$)

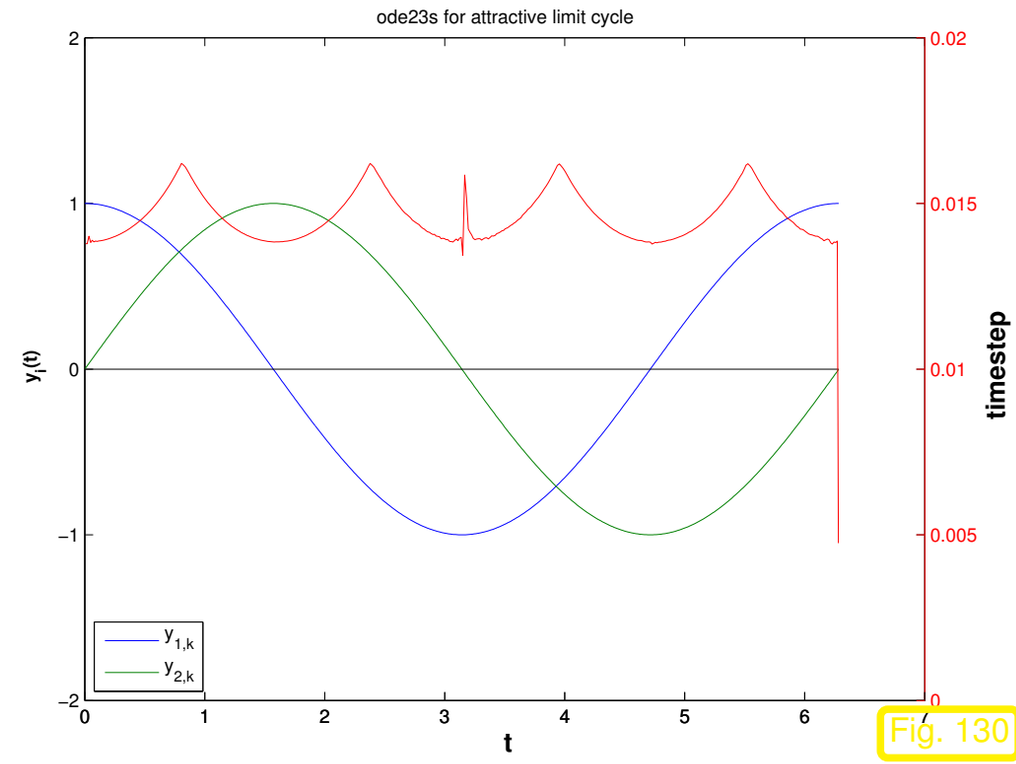
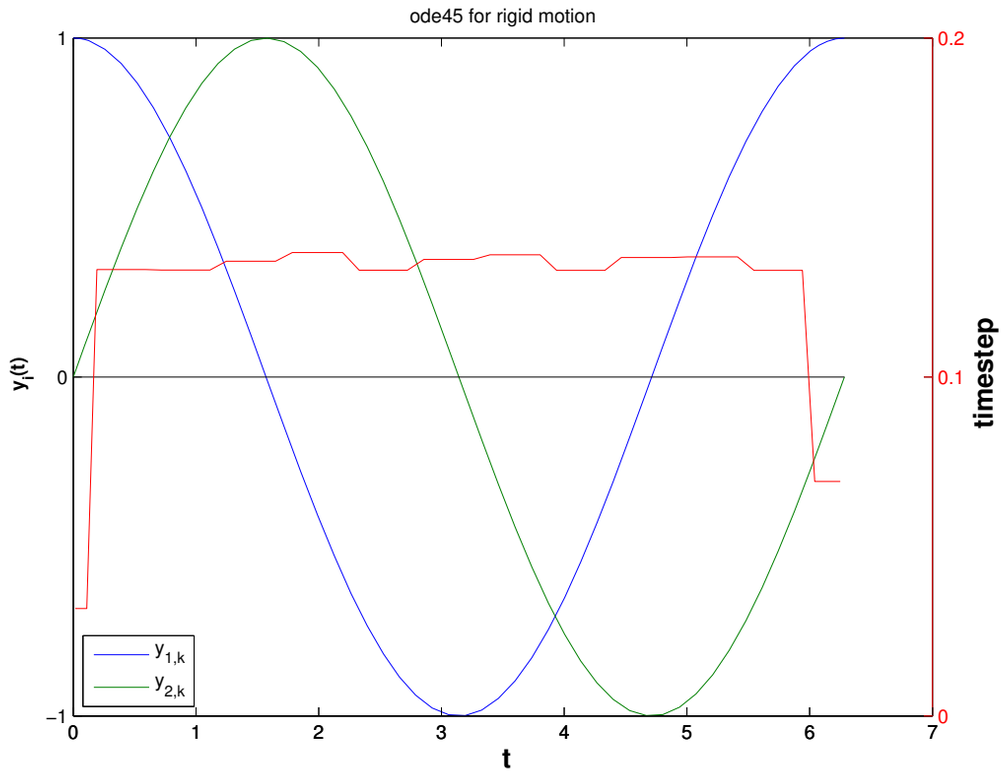


Fig. 130

`ode23s`: 432 steps ($\lambda = 1000$)

P. Grohs

rev 63606, February 17, 2014



◁ $\lambda = 0 \hat{=}$ circular motion, see Ex. 1.4.18

`ode45` achieves good exactness despite (relatively) large stepwidth.

Non-stiff problem!

◇

Adaptive MATLAB-integrators for stiff problems: (Control of stepwidth as in Section 2.6)

```
opts = odeset('abstol',atol,'reltol',rtol,'Jacobian', @J)
[t,y] = ode15s/ode23s(odefun,tspan,y0,opts);
```

Example 3.5.7 (Adaptive semi-implicit RK-OSM for stiff problem). → Ex. 3.5.2, 3.4.1

$$\dot{y}(t) = \lambda y^2(1-y) \;, \quad \lambda = 500 \quad , \quad y(0) = \frac{1}{100} \;.$$

MATLAB-CODE : Semi-implicit OSM for stiff problem

```
lambda = 500; tspan = [0 1]; y0 = 0.01;
fun = @(t,x) lambda*x^2*(1-x);
Jac = @(t,x) lambda*(2*x*(1-x)-x^2);
o = odeset('reltol',0.1,'abstol',0.001,'stats','on','Jacobian',Jac);
[t,y] = ode23s(fun,[0 1],y0,o);
```

Statistic:

```
20 successful steps 4 failed attempts 70 function
evaluations
```

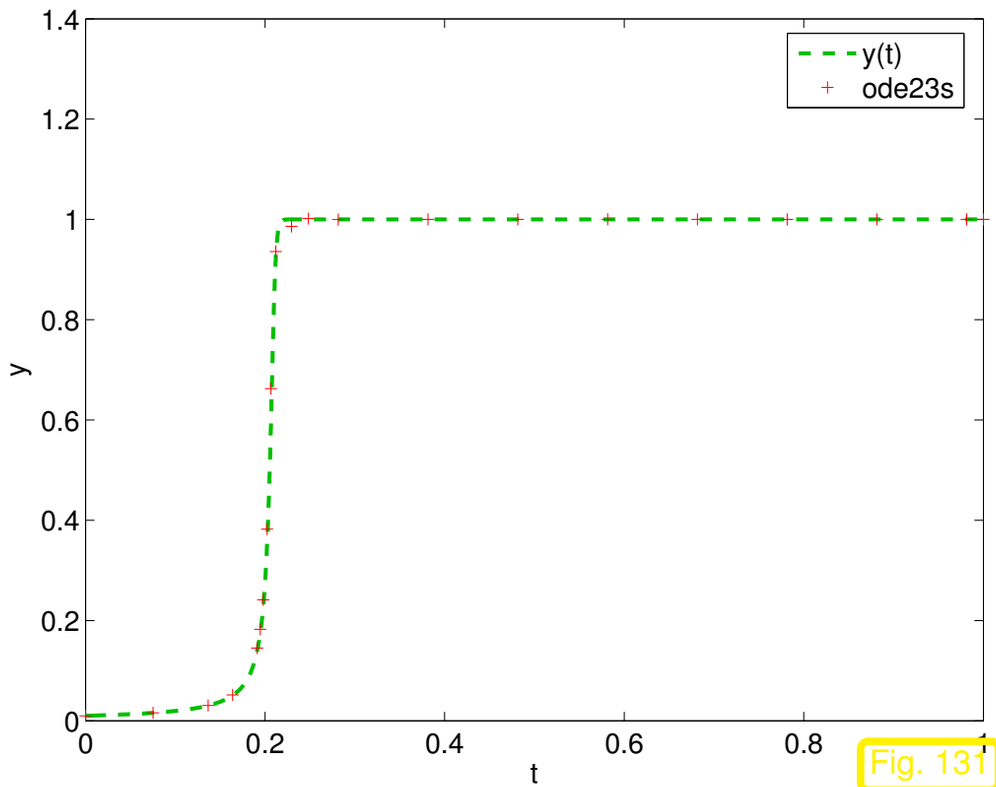


Fig. 131

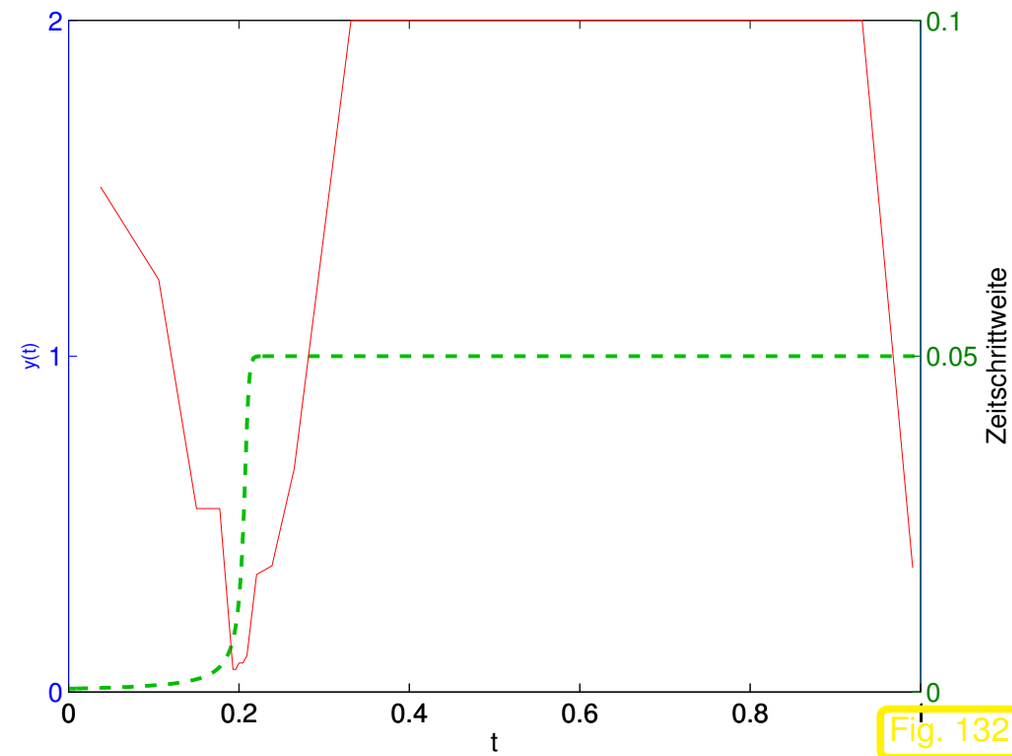


Fig. 132

➤ Efficient method (cf. Ex. 3.5.2): No stepwidth contstrained for $y \approx 1$

3.6 Linear-implicit Runge-Kutta-method [12, Sect. 6.4]

Incremental equation (2.2.3) for s-level implicit RK-OSM

=

Nonlinear system of equation of dimension $s \cdot d$

Example 3.6.1 (Linearising of the incremental equation).

- Initial value problem for logistic differential equation, see Ex. 1.2.1

$$\dot{y} = \lambda y(1-y) \quad , \quad y(0) = 0.1 \quad , \quad \lambda = 5 \, .$$

- Implicit Euler-method (1.4.13) with uniform time step-width $h = 1/n$, $n \in \{5, 8, 11, 17, 25, 38, 57, 85, 128, 192, 288, , 432, 649, 973, 1460, 2189, 3284, 4926, 7389\}$.

& approximately computation of y_{k+1} by one Newton-step with initial value y_k

= semi-implicit Euler-method

- Error measure $\mathrm{err} = \max_{j=1,\dots,n} |y_j - y(t_j)|$

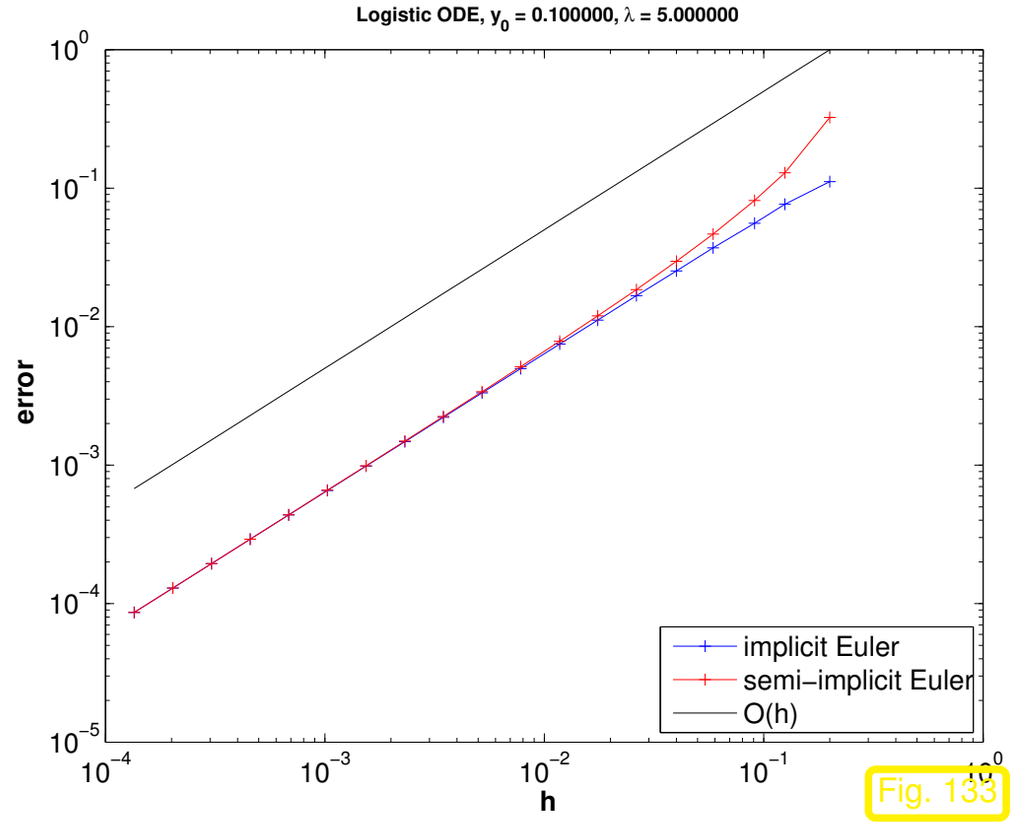


Fig. 133

- Implicit midpoint rule (1.4.19) with uniform time stepwidth $h = 1/n$ (as above)

  & approximately computation of y_{k+1} by one Newton-step with initial value y_k

- Error measure $\mathrm{err} = \max\limits_{j=1,\dots,n} |y_j - y(t_j)|$

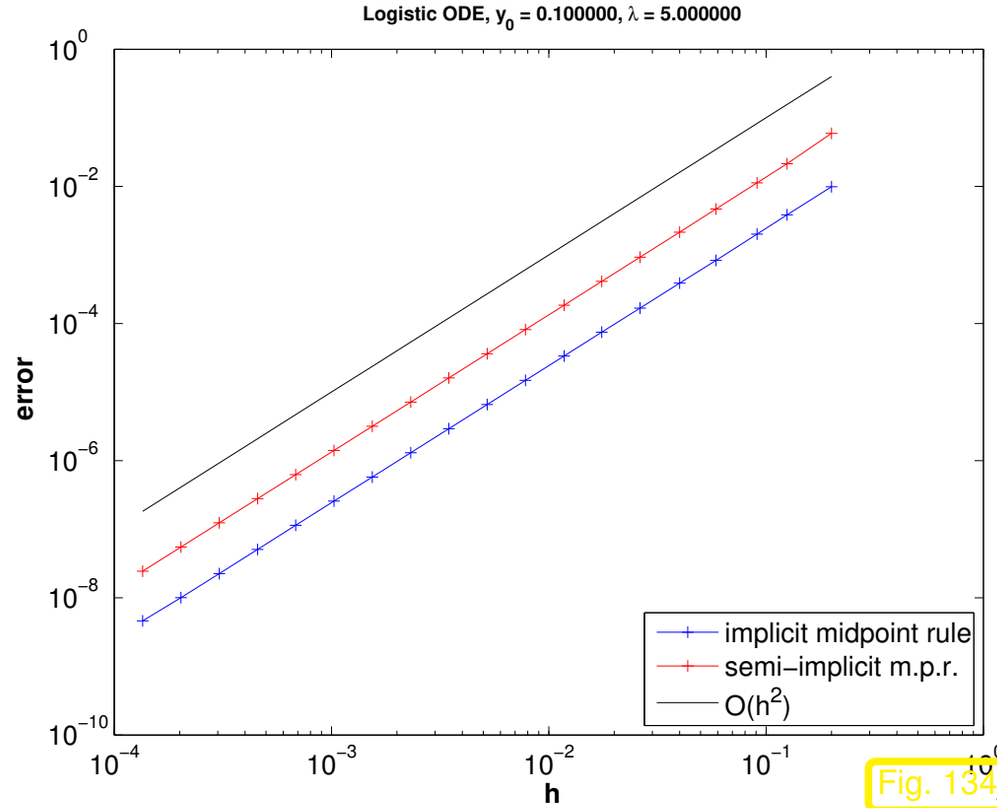
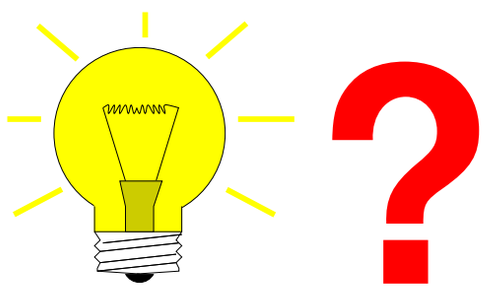


Fig. 134 ◇



Idee: Implicit RK-OSM with linearised incremental equation (2.2.3)

$$\mathbf{k}_i = \mathbf{f}(\mathbf{y}_0) + hD\mathbf{f}(\mathbf{y}_0)\left(\sum_{j=1}^{s} a_{ij}\mathbf{k}_j\right) \quad , \quad i = 1,\dots,s \; . \tag{3.6.2}$$

P. Grohs

rev 63606, February 17, 2014

(3.6.2) $\hat{=}$ LSE of dimension $s \cdot d$: ($s \hat{=}$ number of levels, $\mathfrak{A} \in \mathbb{R}^{s,s} \hat{=}$ coefficient matrix from Butcher-Schema (2.3.6))

$$(\mathbf{I}_{s\cdot d} - h\mathfrak{A} \otimes D\mathbf{f}(\mathbf{y}_0)) \begin{pmatrix} \mathbf{k}_1 \\ \vdots \\ \mathbf{k}_s \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \otimes \mathbf{f}(\mathbf{y}_0) \ ,$$

with Kronecker-product: for $\mathbf{A} \in \mathbb{R}^{m,n}$, $\mathbf{B} \in \mathbb{R}^{k,l}$

$$\mathbf{A} \otimes \mathbf{B} := \begin{pmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & & \vdots \\ a_{m,1}\mathbf{B} & \cdots & a_{m,n}\mathbf{B} \end{pmatrix} \in \mathbb{R}^{m\cdot k, n\cdot l} \ .$$

MATLAB-command `kron(A,B)`.

P. Grohs

rev 63606, February 17, 2014

Linearises with linear ODE ➢ stability function (→ Def. 3.1.6) unchanged

Example 3.6.3 (Implicit RK-OSM with linearised incremental equations).

- Initial value problem for logistic differential equation, see Ex. 1.2.1

$$\dot{y} = \lambda y(1-y) \quad , \quad y(0) = 0.1 \quad , \quad \lambda = 5 \ .$$

- 2-level Radau-OSM, Butcher Schema

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array} , \qquad (3.6.4)$$

Order 3, see Sect. 3.4.

- Increments from linearised equations (3.6.2)
- Error measure $\mathrm{err} = \max_{j=1,\dots,n} |y_j - y(t_j)|$

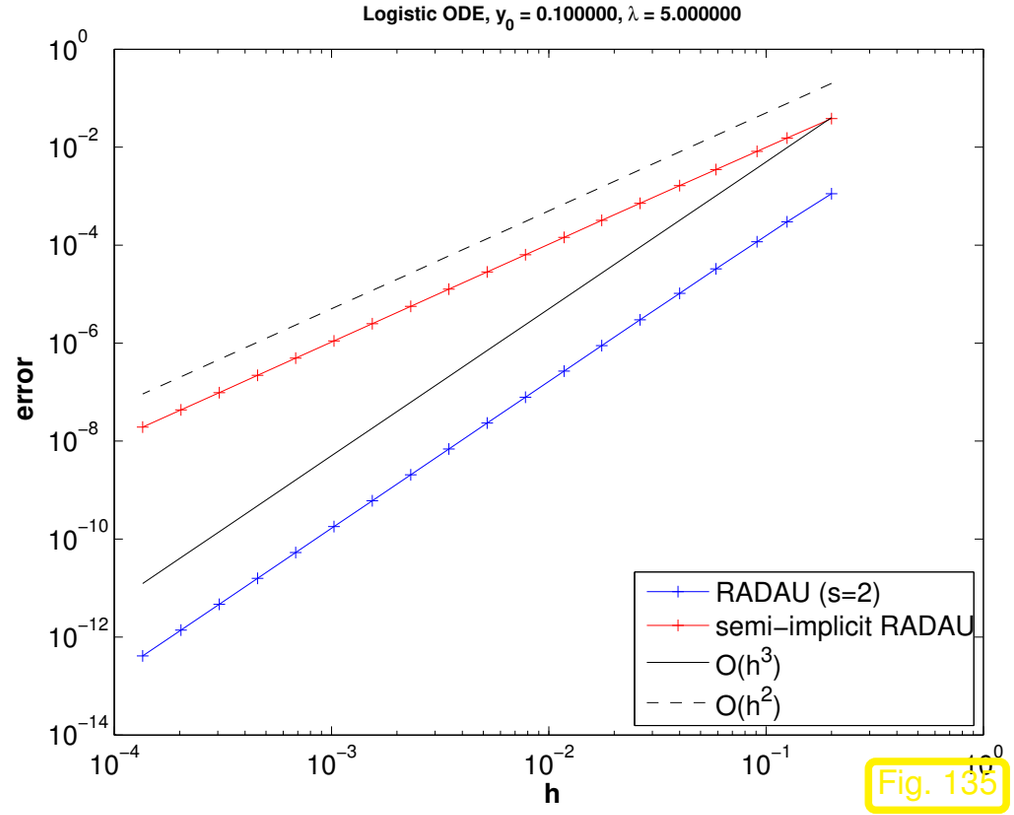


Fig. 135

P. Grohs

rev 63606, February 17, 2014

Order loss by linearisation !

The simple linearisation (3.6.2) leads with implicit RK-OSM to an substantial loss of accuracy and is therefore not an option.

Idea: "“Salvation” of the Order by a better starting approximation for a Newton-step.

We consider:

diagonal-implicit RK-OSM (DIRK)

$\Updownarrow$

$\mathfrak{A}$ lower triangular matrix

($\mathfrak{A}$ regular $\Leftrightarrow a_{jj} \neq 0$)

$$\begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^T \end{array} := \begin{array}{c|cccccc} c_1 & a_{11} & 0 & & \cdots & & 0 \\ c_2 & a_{21} & a_{22} & 0 & & & 0 \\ \vdots & \vdots & & \ddots & \ddots & & \vdots \\ \vdots & \vdots & & & \ddots & \ddots & \vdots \\ \vdots & \vdots & & & & \ddots & 0 \\ c_s & a_{s1} & & \cdots & & & a_{ss} \\ \hline & b_1 & & \cdots & \cdots & & b_s \end{array} . \tag{3.6.5}$$

► Staggered (nonlinear) system of equation for increment (autonomous case)

P. Grohs

rev 63606, February 17, 2014

General incremental equations for s-level DIRK-method:

$$\mathbf{k}_i = \mathbf{f}(\mathbf{y}_0 + h\sum_{j=1}^{i} a_{ij}\mathbf{k}_j) \;, \quad \mathbf{y}_1 = \mathbf{y}_0 + h\sum_{i=1}^{s} b_i\mathbf{k}_i \;.$$

i. Incremental equation: Reformulating as a problem of finding zeros of

$$F(\mathbf{k}_i) := \mathbf{k} - \mathbf{f}(\mathbf{y}_0 + \mathbf{z} + ha_{ii}\mathbf{k}_i) = 0 \;, \quad \mathbf{z} = h\sum_{j=1}^{i-1} a_{ij}\mathbf{k}_j \;.$$

$$DF(\mathbf{k}_i) = \mathbf{I} - D\mathbf{f}(\mathbf{y}_0 + \mathbf{z} + a_{ii}\mathbf{k}_i)ha_{ii} \ .$$

A Newton-step with initial value $\mathbf{k}_i^{(0)}$:

$$\mathbf{k}_i^{(1)} = \mathbf{k}_i^{(0)} - \left(\mathbf{I} - D\mathbf{f}(\mathbf{y}_0 + \mathbf{z} + ha_{ii}\mathbf{k}_i^{(0)})ha_{ii}\right)^{-1} \cdot \left(\mathbf{k}_i^{(0)} - \mathbf{f}(\mathbf{y}_0 + \mathbf{z} + ha_{ii}\mathbf{k}_i^{(0)})\right)$$

Simplification, cf. Rem. 2.3.19: Use Jacobi-Matrix at the position $\mathbf{y}_0$

Newton-method: General Ansatz for starting approximation:

Ansatz starting approximation (for $\mathbf{k}_i$): $$\mathbf{k}_i^{(0)} = \sum_{j=1}^{i-1} \frac{d_{ij}}{a_{ii}}\mathbf{k}_j \ . \tag{3.6.6}$$

$$(\mathbf{I} - ha_{ii}\mathbf{J})\mathbf{k}_i = \mathbf{f}(\mathbf{y}_0 + h\sum_{j=1}^{i-1}(a_{ij} + d_{ij})\mathbf{k}_j) - h\mathbf{J}\sum_{j=1}^{i-1} d_{ij}\mathbf{k}_j \ , \tag{3.6.7}$$

$$\mathbf{J} := D\mathbf{f}\left(\mathbf{y}_0 + \cancel{h\sum_{j=1}^{i-1}(a_{ij} + d_{ij})\mathbf{k}_j}\right) \ . \tag{3.6.8}$$

Simplified Newton-method ("“frozen” Jacobi-Matrix)

As for standard-RK-OSM (→ Def. 2.3.5): $$\mathbf{y}_1 = \mathbf{y}_0 + h\sum_{i=1}^{s} b_i\mathbf{k}_i \ . \tag{3.6.9}$$

Nest step: determine coefficients $\mathbf{a}_{ij}$, $\mathbf{d}_{ij}$ in (3.6.7) (and $\mathbf{b}_i$), such that they satisfy the *order constraints equations*
(analogous of the construction of Runge-Kutta-methods in Sect. 2.3)

► Linear-implicit Runge-Kutta-method (Rosenbrock-Wanner(ROW)-methods)

Example 3.6.10 (Constraint equations for linear-implicit Runge-Kutta-methods of second order).

From the Neumann series for matrices: for $h > 0$ "sufficiently small"

$$(\mathbf{I} - ha_{ii}\mathbf{J})^{-1} = \sum_{k=0}^{\infty}(ha_{ii}\mathbf{J})^k = \mathbf{I} + ha_{ii}\mathbf{J} + O(h^2) \; . \tag{3.6.11}$$

Plugging in (3.6.7) + Taylor-expansion of $\mathbf{f}$ around $\mathbf{y}_0$ + recursive inserting, cf. Ex. 2.3.24:

$$\begin{aligned}\mathbf{k}_i &= \Big(\mathbf{I} + ha_{ii}\mathbf{J} + O(h^2)\Big)\left(\mathbf{f}(\mathbf{y}_0) + h\mathbf{J}\sum_{j=1}^{i-1}(a_{ij}+d_{ij})\mathbf{k}_j + O(h^2) - h\mathbf{J}\sum_{j=1}^{i-1}d_{ij}\mathbf{k}_j\right)\\ &= \mathbf{f}(\mathbf{y}_0) + ha_{ii}\mathbf{J}\mathbf{f}(\mathbf{y}_0) + h\mathbf{J}\mathbf{f}(\mathbf{y}_0)\Big(\sum_{j=1}^{i-1}a_{ij}\Big) + O(h^2)\end{aligned}$$

$$= \mathbf{f}(\mathbf{y}_0) + h\mathbf{J}\mathbf{f}(\mathbf{y}_0)\Big(\sum_{j=1}^{i} a_{ij}\Big) + O(h^2) \ .$$

$$\overset{(3.6.9)}{\Rightarrow} \quad \mathbf{y}_1 = \mathbf{y}_0 + h\Big(\sum_{i=1}^{s} b_i\Big)\mathbf{f}(\mathbf{y}_0) + h^2\Big(\sum_{i=1}^{s} b_i \sum_{j-1}^{i} a_{ij}\Big)\mathbf{J}\mathbf{f}(\mathbf{y}_0) + O(h^3) \ .$$

Where we have used: $\mathbf{J} = D\mathbf{f}(\mathbf{y}_0)$. Then comparison with taylor series (2.3.26) ➢ Constraint equations (2.3.29), (2.3.30) (same as for standard-RK-OSM !).

Example 3.6.12 (Energy conservation with semi-implicit midpoint-rule).

- Hamiltonian ODE (1.2.19) for mathematical pendulum for $0 \leq t \leq T := 50$, initial value $\alpha(0) = \pi/4$, $p(0) = 0$
- Implicit midpoint rule (1.4.19)/semi-implicit midpoint rule ($\rightarrow$ Ex. 3.6.1) on uniform time stepwidth $h = T/300$,
- Observed: Time behavior of the energys $\rightarrow$ Ex. 1.4.17

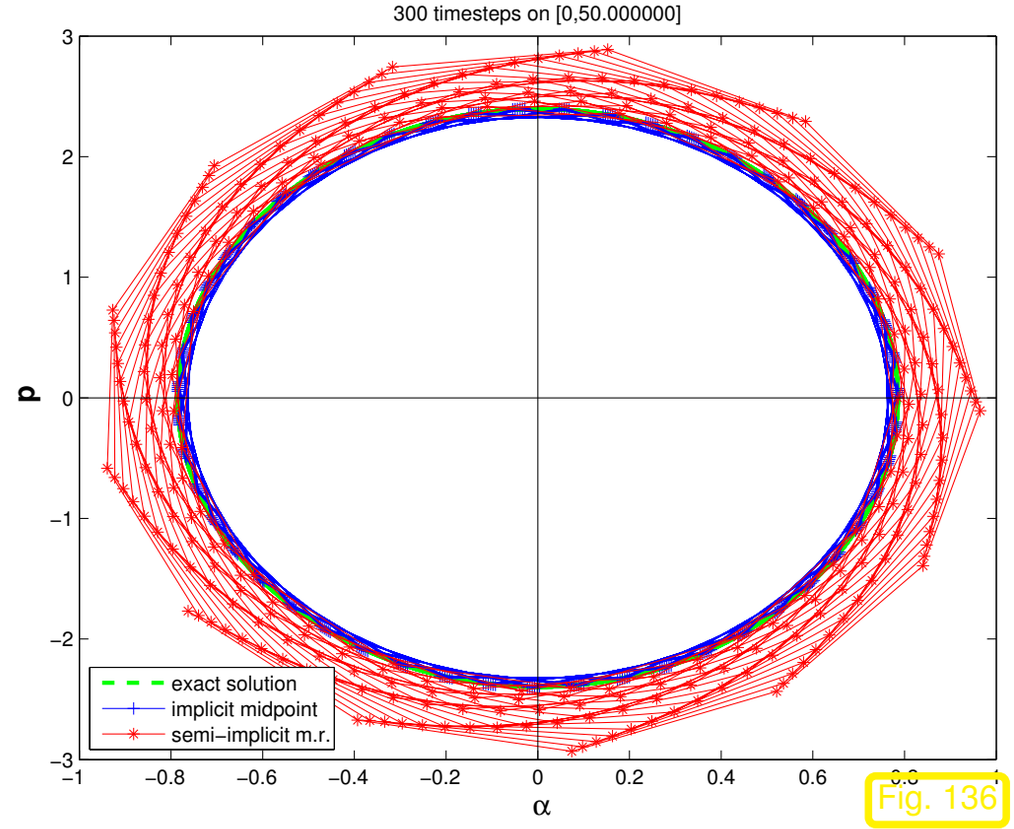


Fig. 136

P. Grohs

rev 63606, February 17, 2014

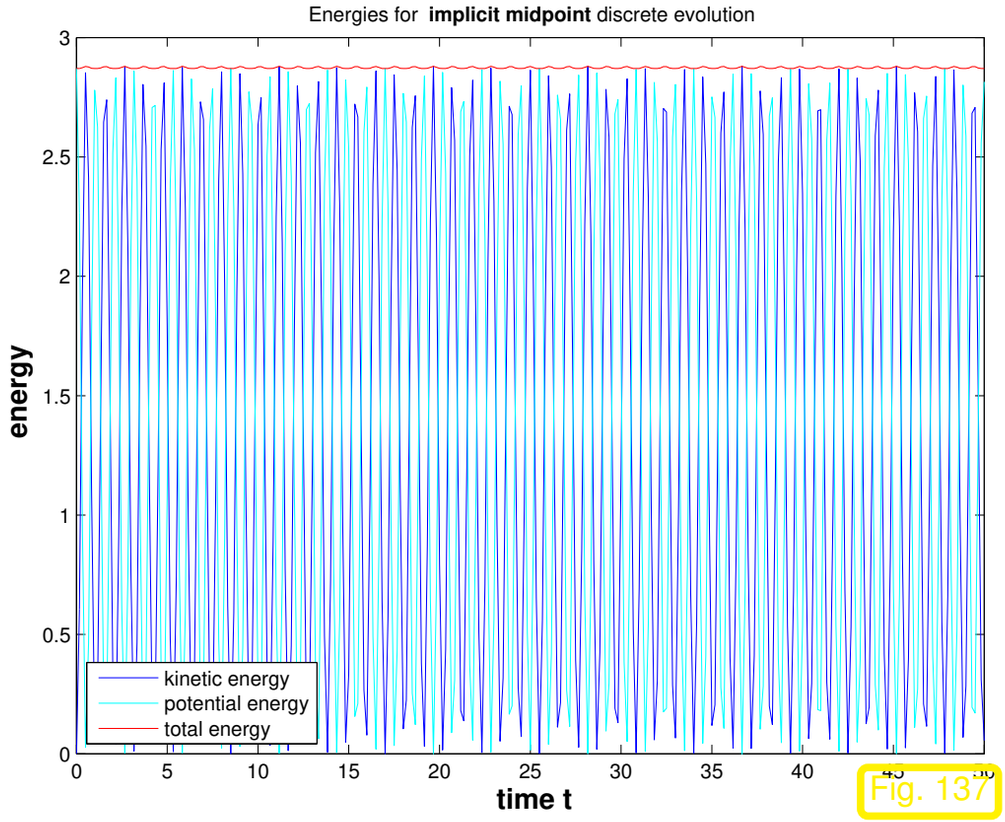


Fig. 137

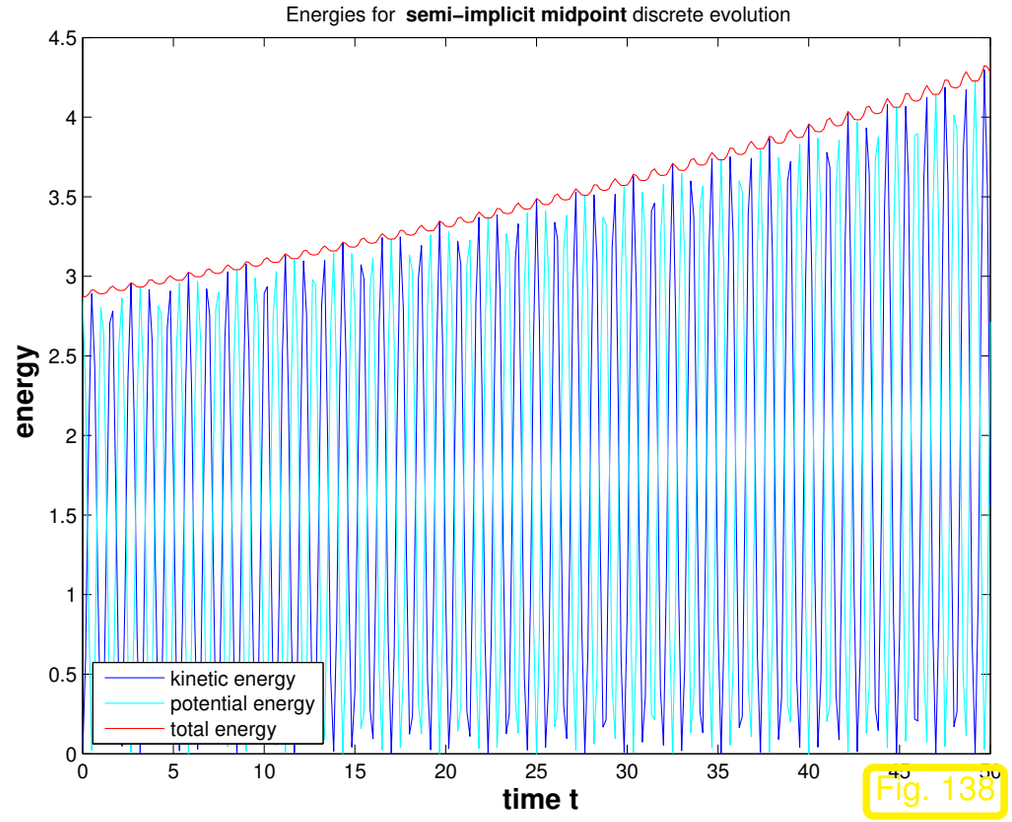


Fig. 138

P. Grohs

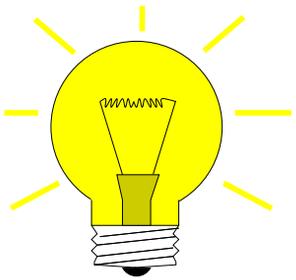
rev 63606, February 17, 2014

Energy drift of semi-implicit midpoint rule

◇

3.7 Exponential Integrators [29, 33, 30]

Note: Autonomous IVP $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\mathbf{f} : D \subset \mathbb{R}^d \mapsto \mathbb{R}^d$ continuously differentiable



Idea: "'Substraction" of the solution of the linearised ODE around $\mathbf{y}_0$

$$\dot{\mathbf{y}} = \mathbf{J}\mathbf{y} + \mathbf{g}(\mathbf{y}) \quad , \quad \mathbf{g}(\mathbf{y}) := \mathbf{f}(\mathbf{y}) - \mathbf{J}\mathbf{y} \; , \tag{3.7.1}$$

with $\boxed{\mathbf{J} := D\mathbf{f}(\mathbf{y}_0)}$

Variation of the constants ($\rightarrow$ Sect. 1.3.2) applied on (3.7.1)

$$\mathbf{y}(h) = \exp(\mathbf{J}h)\mathbf{y}_0 + \int_0^h \exp(\mathbf{J}(h-\tau))\mathbf{g}(\mathbf{y}(\tau))\,\mathrm{d}\tau \; . \tag{3.7.2}$$

convolution integral

$\exp \mathrel{\hat{=}}$ Matrixexponentialfunction, defined by, cf. (1.3.14),

"matrix exponential series": $\exp(\mathbf{M}) = \sum_{k=0}^{\infty} \frac{1}{k!}\mathbf{M}^k$.

! Evaluation of the matrix exponential series is not a stable numerical algorithm (Cancellation !)

Alternatives:

☞ Pade-approximation of $t \mapsto e^t$ (by scaling of the matrix, "‘scaling and squaring”)

☞ Schur-decomposition (see MATLAB-commands `schur`) $\mathbf{M} = \mathbf{Q}^T(\mathbf{D}+\mathbf{U})\mathbf{Q}$ with diagonal matrix $\mathbf{D}$, proper upper triangular matrix $\mathbf{U}$, orthogonal matrix $\mathbf{Q}$. Subsequent evaluation of the cropped matrixexponential series for $\mathbf{D}+\mathbf{U}$ **&** (1.3.15)

MATLAB-function `expm`, algorithm $\rightarrow$ [25]

Numerical quadrature of the convolution integral ➢ discretising of (3.7.2) ➢ OSM

Easiest choice:

$$\int_0^h \exp(\mathbf{J}(h-\tau))g(\mathbf{y}(\tau))\,\mathrm{d}\tau \approx \int_0^h \exp(\mathbf{J}(h-\tau))\,\mathrm{d}\tau \cdot \mathbf{g}(\mathbf{y}_0) = h\varphi(\mathbf{J}h)\cdot\mathbf{g}(\mathbf{y}_0)$$

with $\boxed{\varphi(z) = \frac{\exp(z)-1}{z}}$.

▶ exponential euler-method (on time mesh $\{t_k\}$, $h_k := t_{k+1} - t_k$)

$$\mathbf{y}_{k+1} = \mathbf{y}_k + h_k \varphi(h_k \mathbf{J}) \mathbf{f}(\mathbf{y}_k) \ , \quad k = 0, \dots, N \quad , \quad \mathbf{J} := Df(\mathbf{y}_k) \ . \tag{3.7.3}$$

Remark 3.7.4 (stability domain of the exponential euler-method).

Reminder: Analysis of the *linear* model problems, Sect. 3.1 ➢ stability domain $\mathcal{S}_{\mathbf{\Psi}} \subset \mathbb{C} \rightarrow$ Def. 3.1.4

Note: exponential euler-method is exact for IVP of the ODE $\dot{\mathbf{y}} = \mathbf{A}\mathbf{y} + \mathbf{g}$ with constant $\mathbf{A} \in \mathbb{R}^{d,d}$, $\mathbf{g} \in \mathbb{R}^d$.

▶ (Ideal !) stability domain: $S(z) = \exp(z)$ ➤ (ideal) stability domain $\mathcal{S}_{\mathbf{\Psi}} = \mathbb{C}^-$ △

Example 3.7.5 (Exponential Euler-method).

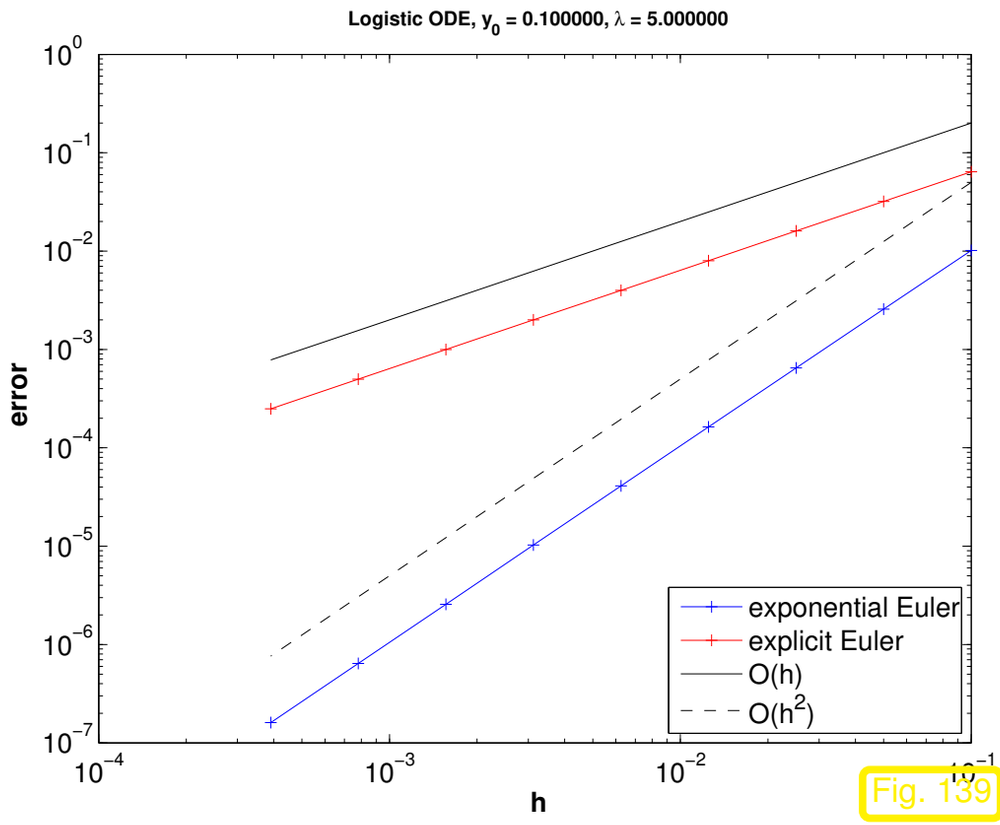


Fig. 139

- Initial value problem for logistic differential equation, $\lambda = 5$, $T = 1$, see Ex. 1.2.1
- Exponential Euler-method (3.7.3) with uniform time-stepwidth h
- measure of error $\quad \mathrm{err} = \max\limits_{j=1,\dots,n} |y_j - y(t_j)|$

▶ algebraic convergence of the 2 Order !

◇

Example 3.7.6 (Exponential Euler-method for stiff IVP).

- Stiff IVP $\rightarrow$ Ex. 3.5.2, 3.4.1, 3.5.7:

$$\dot{y}(t) = \lambda y^2(1-y) \;,$$
$$\lambda = 500 \;, \qquad y(0) = \frac{1}{100} \;.$$

- Exponential Euler-method (3.7.3) with uniform time stepwidth $h = \frac{1}{80}$

▶ Qualitatively correct behavior

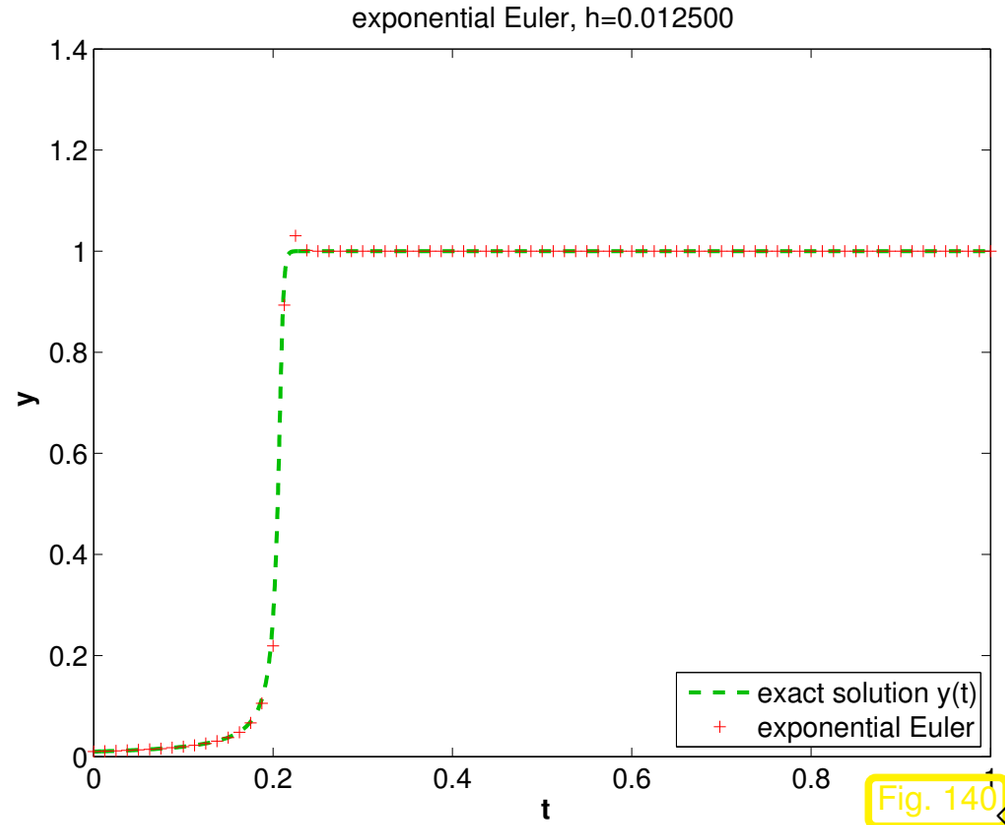


Fig. 140 ◇

P. Grohs

rev 63606, February 17, 2014

Generalization: Exponential Runge-Kutta-method: with $\mathbf{J} := D\mathbf{f}(\mathbf{y}_k)$

Semi-implicit Euler-method	Exponential Euler-method
$\mathbf{y}_{k+1} = \mathbf{y}_k + (\mathbf{I} - h\mathbf{J})^{-1} h\mathbf{f}(\mathbf{y}_k)$	$\mathbf{y}_{k+1} = \mathbf{y}_k + \varphi(h\mathbf{J}) h\mathbf{f}(\mathbf{y}_k)$

☞ Replace: $(\mathbf{I} - \gamma h\mathbf{J})^{-1} \rightarrow \varphi(\gamma h\mathbf{J})$ in linear-implicit RK-OSM (3.6.7)

Definition 3.7.7 (Exponential Runge-Kutta-method)**.**
For $b_i, a_{ij}, d_{ij} \in \mathbb{R}$, $i, j = 1, \dots, s$, $s \in \mathbb{N}$, *defined*

$$\mathbf{k}_i := \varphi(a_{ii} h \mathbf{J})\big(f(\mathbf{u}_i) + h\mathbf{J} \sum_{j=1}^{i-1} d_{ij}\mathbf{k}_j\big) \ , \quad i = 1, \dots, s \quad ,$$

$$\mathbf{u}_i := \mathbf{y}_0 + h \sum_{j=1}^{i-1} (a_{ij} + d_{ij})\mathbf{k}_j \ , \quad i = 1, \dots, s \quad ,$$

$$\mathbf{\Psi}^h \mathbf{y}_0 := \mathbf{y}_0 + h \sum_{i=1}^{s} b_i \mathbf{k}_i \ .$$

a s*-level exponential Runge-Kutta-one-step-method for the autonomous ODE* $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$*. Where* $\mathbf{J} := D\mathbf{f}(\mathbf{y}_0)$*.*

Ase in Sect. 2.3: Desired Consistency order

- ➢ Equations [29]
- ➢ coefficients b_i, a_{ij}, d_{ij}

Additional constraint: exact integration of linear differential equation. $\dot{\mathbf{y}} = \mathbf{A}\mathbf{y} + \mathbf{g}$, $\mathbf{A} \in \mathbb{R}^{d,d}$, $\mathbf{g} \in \mathbb{R}^d$

Remark 3.7.8. Challenge: efficient/exact computation of $\exp(c_i h\mathbf{J})$

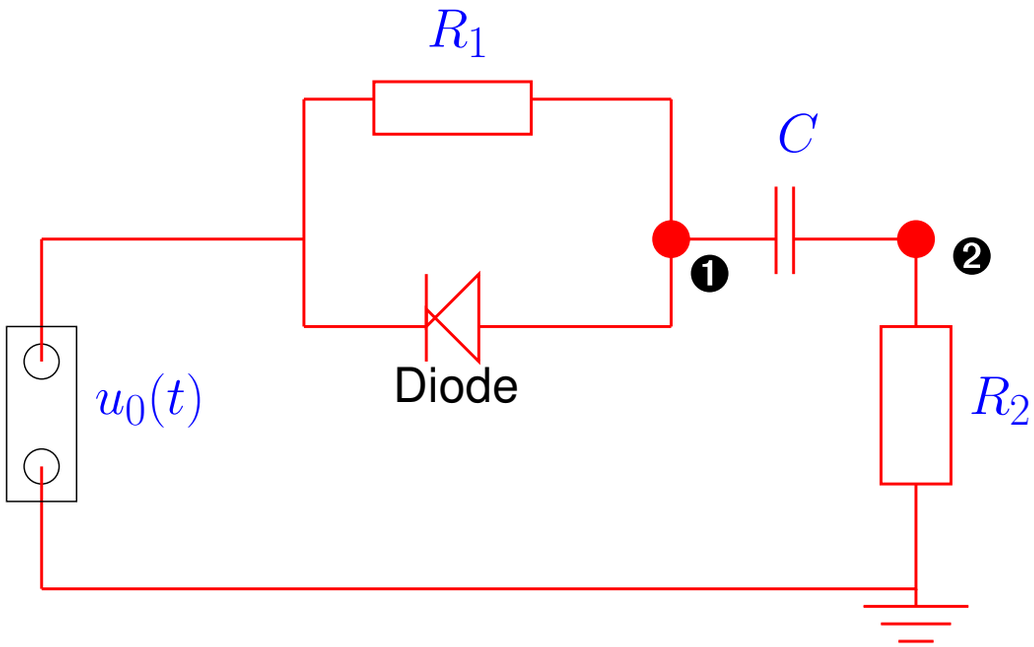
Krylov-subspace method for sparsely populated $\mathbf{J}$ $\rightarrow$ [27]

$\triangle$

3.8 Differential-Algebraic Initial Value Problem

3.8.1 Basic Terms

Example 3.8.1 (Nodes analysis of a circuit). $\rightarrow$ Ex. 3.5.5



Nodes equation (Kirchoff's law):

$$❶: \quad 0 = I_D + I_{R_1} - I_C \ ,$$
$$❷: \quad 0 = I_C - I_{R_2} \ .$$

building panel equations:

$$\begin{aligned} I_D &= I_D(u_1) = I_0(\exp(K(u_0 - u_1)) - 1) \ , \\ I_C &= C(\dot{u}_1 - \dot{u}_2) \ , \\ I_{R_1} &= R_1^{-1}(u_0 - u_1) \ , \\ I_{R_2} &= R_2^{-1} u_2 \ . \end{aligned}$$

Given: Time-independent input voltage $u_0 = u_0(t)$

$$❶ \ \succ \ 0 = I_D(u_1) + R_1^{-1}(u_0 - u_1) - C(\dot{u}_1 - \dot{u}_2) \ ,$$
$$❷ \ \succ \ 0 = C(\dot{u}_1 - \dot{u}_2) - R_2^{-1} u_2 \ .$$

$$\blacktriangleright \quad \begin{pmatrix} C & -C \\ -C & C \end{pmatrix} \begin{pmatrix} \dot{u}_1 \\ \dot{u}_2 \end{pmatrix} = \begin{pmatrix} I_D(u_1) + R_1^{-1}(u_0 - u_1) \\ -R_2^{-1} u_2 \end{pmatrix} \qquad (3.8.2)$$

Singular matrix ! ➥ (3.8.2) is differential-algebraic equation (DAE)

Note:

$$\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} C & -C \\ -C & C \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} C & 0 \\ 0 & 0 \end{pmatrix} .$$

- transformation of (3.8.2): $y_1 := u_1 - u_2$, $y_2 := u_2$

$$\begin{pmatrix} C & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \dot{u}_1 \\ \dot{u}_2 \end{pmatrix} = \begin{pmatrix} C & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \end{pmatrix} = \begin{pmatrix} I_D(u_1) + R_1^{-1}(u_0 - u_1) \\ I_D(u_1) + R_1^{-1}(u_0 - u_1) - R_2^{-1} u_2 \end{pmatrix}$$

$$= \begin{pmatrix} I_D(y_1 + y_2) + R_1^{-1}(u_0 - y_1 - y_2) \\ I_D(y_1 + y_2) + R_1^{-1}(u_0 - y_1 - y_2) - R_2^{-1} y_2 \end{pmatrix} .$$

- *Algebraic constraint*

$$c(y_1, y_2) := I_D(y_1 + y_2) + R_1^{-1}(u_0 - y_1 - y_2) - R_2^{-1} y_2 = 0 .$$

Note: $\forall y_1$: $y_2 \mapsto c(y_1, y_2)$ monotone decreasing, $\lim\limits_{y_2 \to \infty} c(y_1, y_2) = -\infty$, $\lim\limits_{y_2 \to -\infty} c(y_1, y_2) = \infty$

$\Rightarrow$ Constraint is solvable by $y_2 = u_2$: $\exists$ function $G : \mathbb{R} \mapsto \mathbb{R}$ such that $y_2 = G(y_1)$

P. Grohs

rev 63606, February 17, 2014

Plugging in ► $\boxed{\text{ODE for } y_1 \text{ !}}$

$$C\dot{y}_1 = I_D(y_1 + G(y_1)) + R_1^{-1}(u_0 - y_1 - G(y_1)) \ .$$

➢ Existence & Uniqueness of solutions, see Sect. 1.3.1.

◇

Given: Right hand side $\mathbf{f} : D \subset \mathbb{R}^d \mapsto \mathbb{R}^d$,
singular matrix $\mathbf{M} \in \mathbb{R}^{d,d}$

☞ Autonomous (linear) differential-algebraic initial value problem (DAE):

$$\mathbf{M}\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) \quad , \quad \mathbf{y}(0) = \mathbf{y}_0 \ . \tag{3.8.3}$$

Note: (3.8.3) implies algebraic constraint $\boxed{\mathbf{f}(\mathbf{y}(t)) \in \mathrm{Im}(\mathbf{M})}$

(Consistent initial values necessary: $\mathbf{f}(\mathbf{y}_0) \in \mathrm{Im}(\mathbf{M})$!)

✎ Notation: $\mathrm{Im}(\mathbf{M}) := \{\mathbf{M}\mathbf{x} : \mathbf{x} \in \mathbb{R}^d\} \hat{=}$ Image of the matrix $\mathbf{M}$

In Ex. 3.8.1: transformations ➢ Reduction on the special form:

Extension of Def. 1.1.2 ➢ solution concept for (3.8.3)
(difficult: general existence & uniqueness of solutions, see [12, Sect. 2.6])

Given: "'right hand side" $\begin{array}{l}\mathbf{d} : D_1 \times D_2 \subset \mathbb{R}^p \times \mathbb{R}^q \mapsto \mathbb{R}^p, \\ \mathbf{c} : D_1 \times D_2 \subset \mathbb{R}^p \times \mathbb{R}^q \mapsto \mathbb{R}^q\end{array}$ (sufficiently smooth), $p, q \in \mathbb{N}$,

Initial values $\mathbf{u}_0 \in D_1$, $\mathbf{v}_0 \in D_2$.

☞ separated differential-algebraic initial value problem (DAE):

algebraic constraint

$$\begin{array}{l}\dot{\mathbf{u}} = \mathbf{d}(\mathbf{u},\mathbf{v}) \,, \\ \boxed{0 = \mathbf{c}(\mathbf{u},\mathbf{v})}\end{array} \quad , \quad \begin{array}{l}\mathbf{u}(0) = \mathbf{u}_0 \,, \\ \mathbf{v}(0) = \mathbf{v}_0\end{array} \quad , \quad \boxed{\mathbf{c}(\mathbf{u}_0,\mathbf{v}_0) = 0} \,. \tag{3.8.4}$$

consistent initial values necessary!

P. Grohs

rev 63606, February 17, 2014

Remark 3.8.5 (DAE: transformation on separated form).

The form (3.8.3) of a DAE is always in (3.8.4) transformable:

$$\operatorname{rank}(\mathbf{M}) = r \quad \Rightarrow \quad \exists \mathbf{T}, \mathbf{S} \in \mathbb{R}^{d,d} \text{ regulär: } \quad \mathbf{TMS} = \begin{pmatrix} \mathbf{I} & 0 \\ 0 & 0 \end{pmatrix} \ , \quad \mathbf{I} \in \mathbb{R}^{r,r} \ .$$

Application on (3.8.3):

$$\mathbf{M}\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) \quad \Rightarrow \quad \mathbf{TMSS}^{-1}\dot{\mathbf{y}} = \mathbf{Tf}(\mathbf{y}) \quad \overset{\mathbf{z}:=\mathbf{S}^{-1}\mathbf{y}}{\Rightarrow} \quad \begin{pmatrix} \mathbf{I} & 0 \\ 0 & 0 \end{pmatrix} \dot{\mathbf{z}} = \mathbf{Tf}(\mathbf{Sz}) \ .$$

Hence the first r equations of the transformed systems define the differential equation, whereas the remaining $d - r$ play the role of the algebraic constraint.

△

P. Grohs

rev 63606, February 17, 2014

Assumption 3.8.6. *Partial differentiation (Jacobi-matrix)* $D_{\mathbf{v}}\mathbf{c}(\mathbf{u}, \mathbf{v})$ *of the constraints is regular along of the solution curves* $t \mapsto (\mathbf{u}(t), \mathbf{v}(t))^T$.

► Local solvability: $\mathbf{v} = G(\mathbf{u})$: (3.8.4) $\Rightarrow$ $\dot{\mathbf{u}} = \mathbf{d}(\mathbf{u}, G(\mathbf{u})) \hat{=}$ (1.1.13).

Definition 3.8.7 (DAE of Index 1)**.**

Assumption (3.8.6) *satisfies* ➢ *DAE-IVP* (3.8.4) *hat (differentiability)index 1*

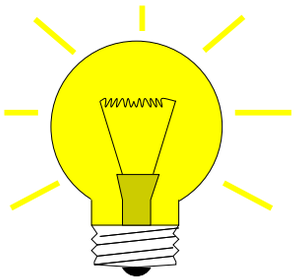
Remark 3.8.8. General discussion of the index term (Index > 1, perturbation index, etc.) for DAE: [23, Kap. VII]

△

3.8.2 Runge-Kutta-method for Index-1-DAEs

Consider: differential-algebraic initial value problem (3.8.4) under the assumption 3.8.6

Idea: Singular perturbation technique (engl. ϵ-*embedding)*



① Consider IVPs for differential equation. $\begin{array}{rcl} \dot{\mathbf{u}} &=& \mathbf{d}(\mathbf{u},\mathbf{v}) \\ \epsilon\dot{\mathbf{v}} &=& \mathbf{c}(\mathbf{u},\mathbf{v}) \end{array}$, $\epsilon > 0$.

② Formulate RK-OSM for $\begin{array}{rcl} \dot{\mathbf{u}} &=& \mathbf{d}(\mathbf{u},\mathbf{v}) \\ \dot{\mathbf{v}} &=& \frac{1}{\epsilon}\mathbf{c}(\mathbf{u},\mathbf{v}) \end{array}$, $\epsilon > 0$.

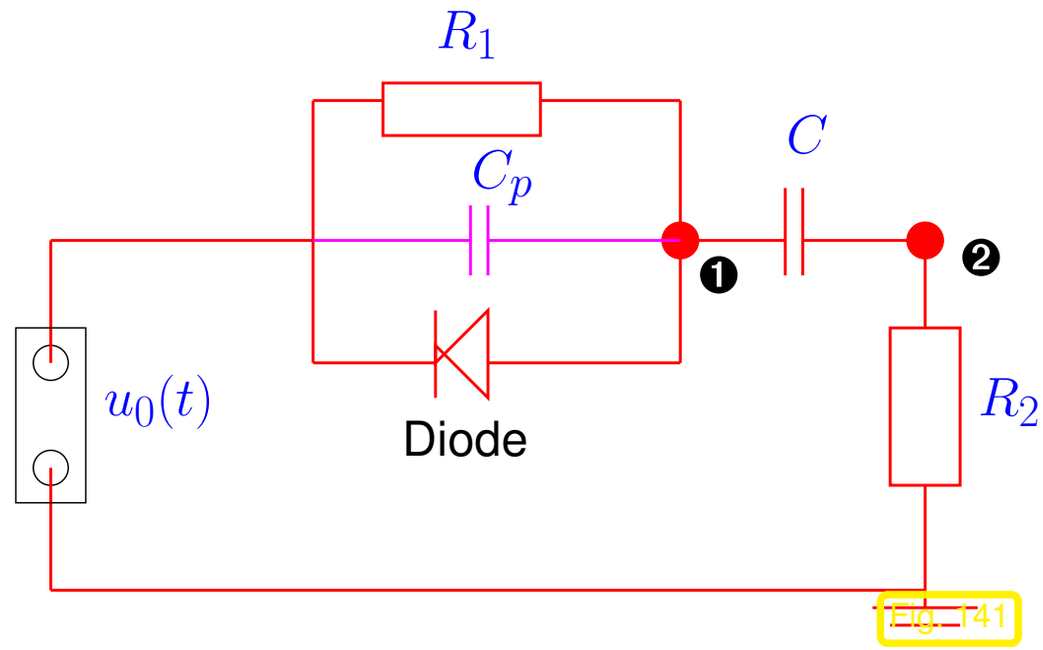
③ Does method still make sense for $\epsilon = 0$ **?** If yes $\rightarrow$ ☺

[Beitrag von J. Mitrovic, FS 2011]

In mathematical models ϵ is often the “insignificant” components.

Example 3.8.9 (Singular disturbed circuit equations).

Circuits form Ex. 3.8.1 with parasitic capacity (flown by electric current I_p)



Nodes equations (Kirchoff's rule):

❶: $0 = I_D + I_{R_1} + I_p - I_C$,

❷: $0 = I_C - I_{R_2}$.

Additional building panel equation:

$$I_p = C_p(\dot{u}_0 - \dot{u}_1) \ .$$

$$\begin{pmatrix} C+C_p & -C \\ -C & C \end{pmatrix} \begin{pmatrix} \dot{u}_1 \\ \dot{u}_2 \end{pmatrix} = \begin{pmatrix} I_D(u_1) + R_1^{-1}(u_0 - u_1) + C_p\dot{u}_0 \\ -R_2^{-1}u_2 \end{pmatrix} \tag{3.8.10}$$

Regular matrix for $C_p > 0$

P. Grohs

rev 63606, February 17, 2014

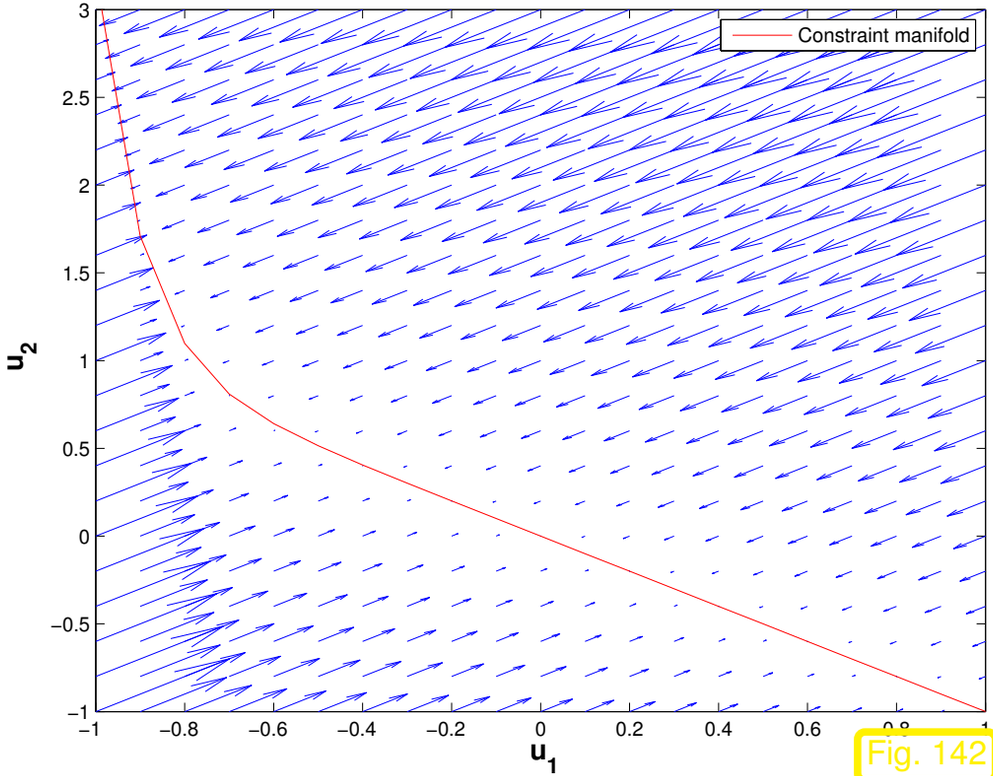


Fig. 142

◁ Direction field of the singular disturbed circuit equations for $u_0(t) \equiv 0$

(Scaled sizes $R = 1$, $C = 1$, $I_0 = 10^{-4}$, $K = 10$, $C_p = 10^{-3}$)

P. Grohs

rev 63606, February 17, 2014

From the directional field we infer: fast relaxation in direction of the manifold described by the algebraic constraint of the DAE (3.8.2)

$$u_2 = R_2(I_D(u_1) + R_1^{-1})(u_0 - u_1) \ .$$

☞ Stiffness of the singular disturbed problems, see Ex. 3.5.6.

Quantitative Analysis: Consider the case $u_0(t) \equiv 0$ ➢ Stationary point: $u_1 = 0$, $u_2 = 0$

Jacobi-matrix in the stationary point

$$Df(0) = C_p^{-1} \begin{pmatrix} 1 & 1 \\ 1 & 1 + \frac{C_p}{C} \end{pmatrix} \begin{pmatrix} -I_0 K - R_1^{-1} & 0 \\ 0 & -R_2^{-1} \end{pmatrix}$$

➢ $C_p \to 0 \quad \Rightarrow \quad \lambda_{\min}(Df(0)) \to -\infty$

DAEs = "'∞-stiff initial value problem"

◇

P. Grohs

rev 63606, February 17, 2014

Formal computation: Singulare perturbation technique for Runge-Kutta-one-step-method with Butcher-Scheme $\begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^T \end{array}$

Def. 2.3.5: s-level Runge-Kutta-step for $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, echolon form, see Rem. 2.3.7:

$$\begin{aligned} \mathbf{k}_i &= \mathbf{f}(\mathbf{y}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j) \\ \mathbf{y}_1 &= \mathbf{y}_0 + h\sum_{i=1}^{s} b_i\mathbf{k}_i \end{aligned} \quad \overset{\mathbf{k}_i=\mathbf{f}(\mathbf{g}_i)}{\Leftrightarrow} \quad \begin{aligned} \mathbf{g}_i &= \mathbf{y}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{f}(\mathbf{g}_j) \\ \mathbf{y}_1 &= \mathbf{y}_0 + h\sum_{i=1}^{s} b_i\mathbf{f}(\mathbf{g}_i) \end{aligned} \quad , \quad i = 1,\dots,s\,.$$

$$\begin{cases} \dot{\mathbf{u}} = \mathbf{d}(\mathbf{u},\mathbf{v}) \\ \dot{\mathbf{v}} = \frac{1}{\epsilon}\mathbf{c}(\mathbf{u},\mathbf{v}) \end{cases} \quad \blacktriangleright \quad \begin{cases} \mathbf{g}_i^u = \mathbf{u}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{d}(\mathbf{g}_j^u,\mathbf{g}_j^v) \\ \epsilon\mathbf{g}_i^v = \epsilon\mathbf{v}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{c}(\mathbf{g}_j^u,\mathbf{g}_j^v) \end{cases} , i = 1,\dots,s$$

$$\begin{cases} \mathbf{u}_1 = \mathbf{u}_0 + h\sum_{i=1}^{s} b_i\mathbf{d}(\mathbf{g}_i^u,\mathbf{g}_i^v) \\ \mathbf{v}_1 = \mathbf{v}_0 + \frac{1}{\epsilon}h\sum_{i=1}^{s} b_i\mathbf{c}(\mathbf{g}_i^u,\mathbf{g}_i^v) \end{cases}$$

$$\epsilon \to 0 \quad \Rightarrow \quad \sum_{j=1}^{s} a_{ij}\mathbf{c}(\mathbf{g}_j^u,\mathbf{g}_j^v) = 0,\ i = 1,\dots,s \quad \Rightarrow \quad \boxed{\mathbf{c}(\mathbf{g}_j^u,\mathbf{g}_j^v) = 0}$$

Assumption: coefficient matrix $\mathfrak{A} := (a_{ij})_{i,j=1}^{s}$ regular

P. Grohs

rev 63606, February 17, 2014

① If constraint by $\mathbf{v}$ solvable ($\rightarrow$ Index 1, Def. 3.8.7)

$$\mathbf{c}(\mathbf{u}_1, \mathbf{v}_1) = 0 \quad \Rightarrow \quad \mathbf{v}_1 = G(\mathbf{u}_1) \ .$$

This can replace the step $\mathbf{v}_0 \mapsto \mathbf{v}_1$.

② In the general case, formally, with $\mathfrak{A}^{-1} = (\check{a}_{ij})_{i,j=1}^s$

$$\frac{1}{\epsilon} h\mathbf{c}(\mathbf{g}_i^u, \mathbf{g}_i^v) = \sum_{j=1}^s \check{a}_{ij}(\mathbf{g}_j^v - \mathbf{v}_0)$$

$$\Rightarrow \qquad \mathbf{v}_1 = \mathbf{v}_0 + \sum_{i=1}^s b_i \sum_{j=1}^s \check{a}_{ij}(\mathbf{g}_j^v - \mathbf{v}_0) = \underbrace{(1 - \mathbf{b}^T \mathfrak{A}^{-1} \mathbf{1})}_{= S(-\infty)} \mathbf{v}_0 + \sum_{j=1}^s \Big(\sum_{i=1}^s b_i \check{a}_{ij}\Big) \mathbf{g}_j^v \ .$$

From formula (3.4.4) for stability function S

Note: $\mathbf{c}(\mathbf{u}_1, \mathbf{v}_1) = 0$ is not guaranteed! (➢ Method drifts away from the manifold defined by the algebraic constraint.)

Special case: If RK-OSM stiffly accurate, also $\mathbf{b}^T = \mathbf{a}_{\cdot,s}^T$ (row of $\mathfrak{A}$), cf. sufficient condition (3.4.5) for L-stability

$$\Rightarrow \quad \mathbf{v}_1 = \mathbf{g}_s^v \quad \Rightarrow \quad \mathbf{c}(\mathbf{u}_1, \mathbf{v}_1) = 0$$

Remark 3.8.11 (RK-OSM and elimination of the DAE-constraints).

Index-1 DAE ($\rightarrow$ Def. 3.8.7) $\leftrightarrow$ ODE $\dot{\mathbf{u}} = \mathbf{d}(\mathbf{u}, G(\mathbf{u}))$

stiffly accurate RK-OSM according to (3.4.5) for this ODE:

$$\mathbf{g}_i^u = \mathbf{u}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{d}(\mathbf{g}_j^u, G(\mathbf{g}_j^u)) \;, \quad i = 1,\dots,s \quad , \quad \mathbf{u}_1 = \mathbf{g}_s^u$$

$$\Updownarrow$$

$$\begin{cases} \mathbf{g}_i^u = \mathbf{u}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{d}(\mathbf{g}_j^u, \mathbf{g}_j^v) \\ 0 = \mathbf{c}(\mathbf{g}_i^u, \mathbf{g}_i^v) \end{cases} \;, \quad i = 1,\dots,s \quad , \quad \mathbf{u}_1 = \mathbf{g}_s^u \;.$$

☞ A “commutative diagram” (for sufficiently small h)

Stiffly accurate RK-OSM for $\begin{array}{rcl} \dot{\mathbf{u}} &=& \mathbf{d}(\mathbf{u},\mathbf{v}) \\ 0 &=& \mathbf{c}(\mathbf{u},\mathbf{v}) \end{array}$ **=** Stiffly accurate RK-OSM for $\dot{\mathbf{u}} = \mathbf{d}(\mathbf{u}, G(\mathbf{u}))$

△

Which Runge-Kutta-one-step method ($\rightarrow$ Sect. 2.3)
are suitable for the singular perturbation technique **?**

- Necessary (for the solvability of the incremental equation, Def. 2.3.5): implicit method
- DAE "'∞-stiff" ➢ Necessary: A-stability $\rightarrow$ Def. 3.2.16,
  Desirable: L-stability $\rightarrow$ Def. 3.4.3

Again: Insights by model problem analysis, cf. Sect. 3.1:

Model problem: $\begin{aligned} \dot{u} &= v f(u) \,, \\ 0 &= 1 - v \end{aligned}$ ▶ $\begin{aligned} \dot{u} &= v f(u) \,, \\ \epsilon \dot{v} &= 1 - v \end{aligned}$ $\quad 0 < \epsilon \ll 1 \,.$

Singular disturbed differential equation

Heuristic: OSM suitable for model problem $\leftrightarrow$ OSM suitable for singular disturbed problem $\forall \epsilon \approx 0$

▶ OSM for IVP $\dot{v} = \epsilon^{-1}(1-v)$, $v(0) = 1$, $0 < \epsilon \ll 1$, must yield the sequence $\{v_k\}_{k=0}^{\infty}$ with $\lim\limits_{k\to\infty} v_k = 1$!

▶ Desirable: $\boxed{S(-\infty) = 0}$ for the stability function ($\rightarrow$ Thm. 3.1.6) $S(z)$ of the OSM.

Theorem 3.8.12 (convergence implicit RK-OSM for Index-1-DAEs). → *[23, Thm. 1.1, Sect. VI.1]*

Let $\mathbf{d}$, $\mathbf{c}$ *be sufficiently smooth, the initial value problem* (3.8.4) *uniquely solvable and assumption 3.8.6 satisfied (→ Index-1-DAE, see Def. 3.8.7). The* s*-level Runge-Kutte-one-step-method with Butcher-tableau* $\begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^T \end{array}$*, see* (2.3.6)*, let stiffly-accurate, i.e.* $\mathfrak{A}$ *is regular and* $b_i = a_{s,i}$*,* $i = 1, \dots, s$*, and let be consistent of order* p*.*

Then the method applied on (3.8.4) *is well-defined for sufficiently small time step-width* h *and we have*

$$\|\mathbf{u}_k - \mathbf{u}(t_k)\| = O(h^p) \quad , \quad \|\mathbf{v}_k - \mathbf{v}(t_k)\| = O(h^p) \ ,$$

on each finite integration time-interval $[0, T]$*.*

Sketch of the proof: Solve (3.8.4) for the algebraic variable $\mathbf{v}$ and insert it in ➢ ODE

Application of the RK-OSM on the resulting ODE yields exactly the same discrete evolution for $\mathbf{u}$ as the method for the DAE ("commutative diagram"), cf. Rem. 3.8.11.

Numeric integration of Index-1-DAEs with Radau-OSM → Sect. 3.4

Remark 3.8.13. Radau-OSM also appropriate for Index-1-DAEs (!) in the form (3.8.3) $\mathbf{M}\dot{\mathbf{y}} = f(\mathbf{y})$ △

Remark 3.8.14 (MATLAB-integrators for Index-1-DAEs).

MATLAB-CODE : Lösung einer Index-1-DAE

```
f = @(t,y) [ ...  ];
J = @(t,y) [ ...  ];
M = @(t,y) [ ...  ];
opts = odeset('Mass',M,'Jacobian',J);
[t,y] = ode15s(f,tspan,y0,opts);
```

MATLAB-Code for the solution of general DAEs

$$\mathbf{M}(t,\mathbf{y})\dot{\mathbf{y}} = f(t,\mathbf{y}) \ .$$

function f

Jacobi-Matrix $D_{\mathbf{y}} f$

Matrix(function) $\mathbf{M}(t,\mathbf{y})$

Alternative integrator: ode23t (same call)

(Note: All MATLAB DAE-integrators use adaptive control of stepwidth, cf. Sect. 2.6) △

Example 3.8.15 (Solution of the circuit-DAEs with MATLAB).

- DAE (3.8.2) with $C = 1$, $R_2 = 1$, $R_1 = 1000$, $K = 10$, $I_0 = 10^{-1}$
- MATLAB-integrators ode23t, ode15s, default tolerances

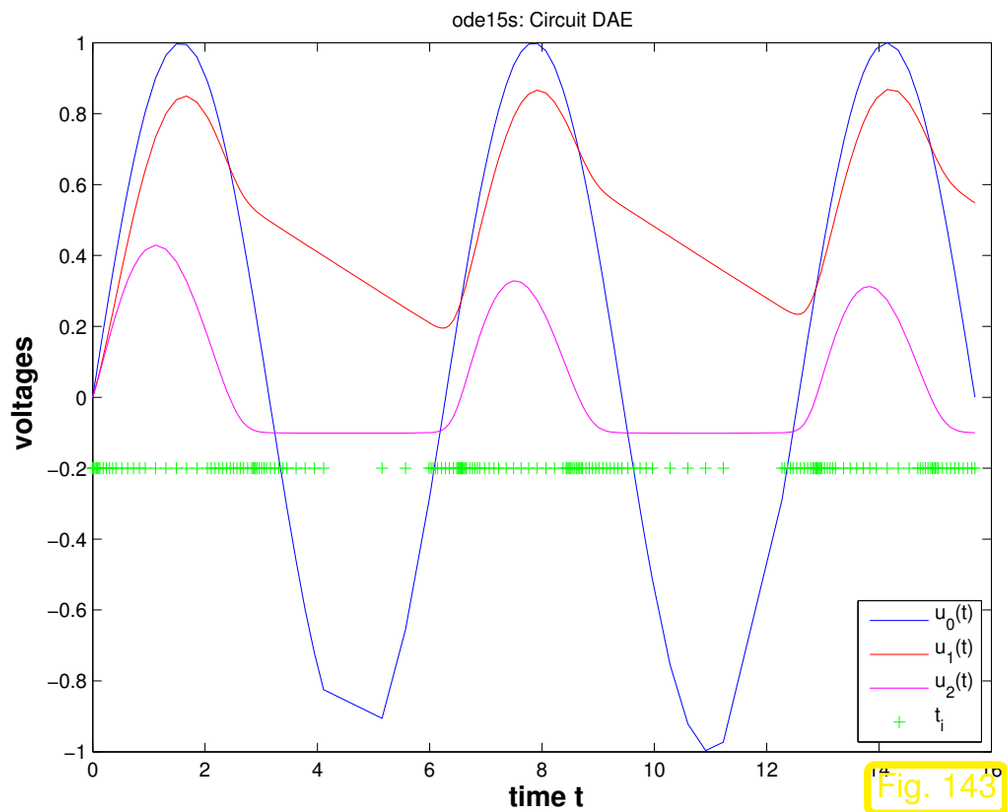


Fig. 143

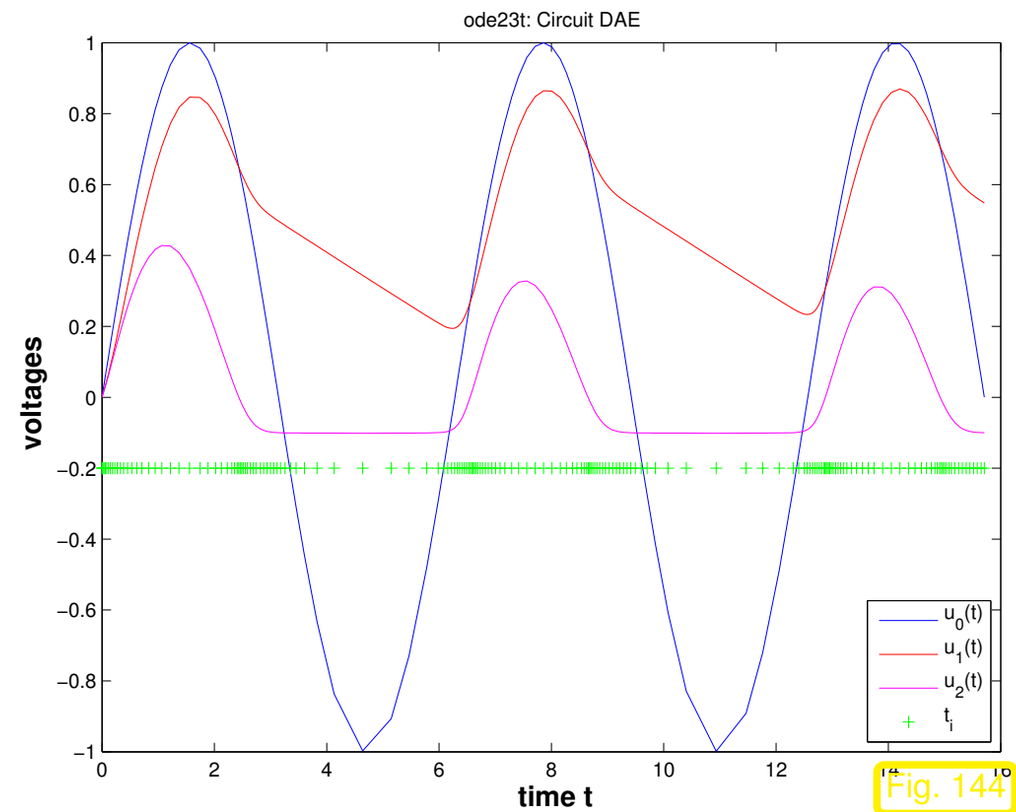


Fig. 144 ◇

3.8.3 DAEs with higher index

Example 3.8.16 (Pendulum equation in descriptive form). $\rightarrow$ Ex. 1.2.17

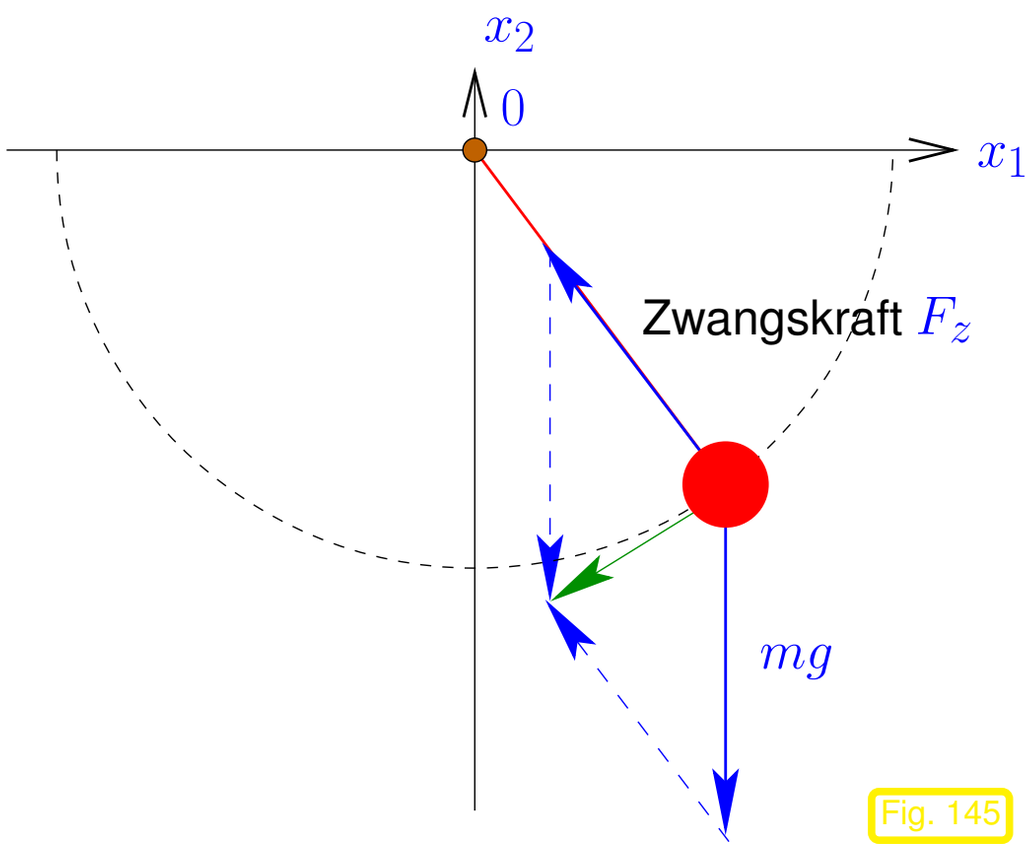


Fig. 145

Mathematical pendulum (Fixed in 0)

Constraining force holds the mass of the pendulum on the circular path

$$F_z(\mathbf{x}) - mg\begin{pmatrix}0\\1\end{pmatrix} \perp \mathbf{x}\,. \tag{3.8.17}$$

Constraining force in direction of the stick:

$$F_z(\mathbf{x}) = -\lambda m\mathbf{x}\,. \tag{3.8.18}$$

($\mathbf{x} = (x_1, x_2)^T \hat{=}$ position of the mass)

▶ Equations of motion in descriptive form: ($\leftrightarrow$ minimal coordinates in Ex. 1.2.17)

$$\ddot{\mathbf{x}} = \lambda\mathbf{x} - g\begin{pmatrix}0\\1\end{pmatrix} \quad , \quad \boxed{x_1^2 + x_2^2 = l^2}\,. \tag{3.8.19}$$

Lagrange-multiplicators ➢ Constraining force

constraining force

(3.8.19) $\hat{=}$ second order ➢ Transformation in separated DAE (3.8.4) of first order:

$$(3.8.19) \quad ➢ \quad m \begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{p}_1 \\ \dot{p}_2 \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ -\lambda x_1 \\ -\lambda x_2 - g \end{pmatrix} \quad , \quad x_1^2 + x_2^2 = l^2 \; . \tag{3.8.20}$$

▶ (3.8.20) is DAE of Index > 1 **!**

differentiation of the constraining forces ▶ $x_1 p_1 + x_2 p_2 = 0 \; ,$ (3.8.21)

additional differentiation ▶ $(p_1^2 + p_2^2) - \lambda(x_1^2 + x_2^2) - g x_2 = 0 \; .$ (3.8.22)

P. Grohs

rev 63606, February 17, 2014

- (3.8.21), (3.8.22) $\hat{=}$ hidden constraints (to be fulfilled by the initial values!)
- Not before the second differentiation of the constraint we can the resultant constraint (3.8.22) solve for λ ➥ (3.8.20) has index 3

◇

Remark 3.8.23 (Hamiltonian equation of motion with constraints).

Consider: Mechanical system with *Hamiltonian-functions* (→ Def. 1.2.20) $H = H(\mathbf{p}, \mathbf{q})$
($\mathbf{q} \in \mathbb{R}^n \hat{=}$ configuration variable, $\mathbf{p} \in \mathbb{R}^n \hat{=}$ impuls variable, see Sect. 1.2.4)

Constraining force for configurations: $\boxed{\mathbf{c}(\mathbf{q}(t)) = 0 \; \forall t \geq 0, \; \mathbf{c} : \mathbb{R}^n \mapsto \mathbb{R}^m, \; m < n}$

► Hamiltonian equations of motion with constraints

$$\begin{aligned} \dot{\mathbf{q}} &= \frac{\partial H}{\partial \mathbf{p}}(\mathbf{p}, \mathbf{q}) \\ \dot{\mathbf{p}} &= -\frac{\partial H}{\partial \mathbf{q}}(\mathbf{p}, \mathbf{q}) - D\mathbf{c}(\mathbf{q})^T \boldsymbol{\lambda} \\ 0 &= \mathbf{c}(\mathbf{q}) \; . \end{aligned} \tag{3.8.24}$$

Lagrange-multiplier $\boldsymbol{\lambda} : \mathbb{R} \mapsto \mathbb{R}^m$

If $m = 1$ ➢ $c(\mathbf{q}) = 0$ describes $n-1$-dimensional manifold in $\mathbb{R}^n$

$\mathbf{grad}\, c(\mathbf{q}) \hat{=}$ normal vector on this manifold

$\lambda\, \mathbf{grad}\, c(\mathbf{q}) \hat{=}$ constraining force *orthogonal* to the manifold, cf. (3.8.17)

Frequent Special case: with $\mathbf{M} \hat{=}$ **s.p.d.** mass matrix, $U \hat{=}$ potential,

$$H(\mathbf{p},\mathbf{q}) = \underbrace{\frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p}}_{\text{kinetic energy}} + \underbrace{U(\mathbf{q})}_{\text{potential energy}} \tag{3.8.25}$$

(3.8.24) in this special case:

$$\dot{\mathbf{q}} = \mathbf{M}^{-1}\mathbf{p}\,, \quad \dot{\mathbf{p}} = -\,\mathbf{grad}\,U(\mathbf{q}) - D\mathbf{c}(\mathbf{q})^T\boldsymbol{\lambda}\,, \quad \mathbf{c}(\mathbf{q}) = 0\,,$$

$$\blacktriangleright \qquad \mathbf{M}\ddot{\mathbf{q}} = -\,\mathbf{grad}\,U(\mathbf{q}) - D\mathbf{c}(\mathbf{q})^T\boldsymbol{\lambda}\,, \quad \mathbf{c}(\mathbf{q}) = 0\,.$$

of this form is e.g. the pendulum equation (3.8.20).

Now: Differentiation of the constraint force $\mathbf{c}(\mathbf{q}) = 0$ by time **+** application of the product rule & chain rule **+** insert of the differential equation from (3.8.24).

➢ hidden constraint $\leftrightarrow$ (3.8.21), (3.8.22)

$$0 = D\mathbf{c}(\mathbf{q})\frac{\partial H}{\partial \mathbf{p}}(\mathbf{p},\mathbf{q}) \;, \tag{3.8.26}$$

$$\begin{aligned} 0 = {} & D^2\mathbf{c}(\mathbf{q})(\frac{\partial H}{\partial \mathbf{p}}(\mathbf{p},\mathbf{q}),\frac{\partial H}{\partial \mathbf{p}}(\mathbf{p},\mathbf{q})) + D\mathbf{c}(\mathbf{q})\frac{\partial^2 H}{\partial \mathbf{p}\partial \mathbf{q}}(\mathbf{p},\mathbf{q})\frac{\partial H}{\partial \mathbf{p}}(\mathbf{p},\mathbf{q}) \\ & - D\mathbf{c}(\mathbf{q})\frac{\partial^2 H}{\partial^2 \mathbf{p}}(\mathbf{p},\mathbf{q})(\frac{\partial H}{\partial \mathbf{q}}(\mathbf{p},\mathbf{q}) + D\mathbf{c}(\mathbf{q})^T\boldsymbol{\lambda}) \;. \end{aligned} \tag{3.8.27}$$

For H of the form (3.8.25): $\dfrac{\partial^2 H}{\partial \mathbf{p}^2} = \mathbf{M}^{-1}$ (invertible matrix)

➢ (3.8.27) solvable for $\boldsymbol{\lambda}$ if $D\mathbf{c}(\mathbf{q})^T$ injective $\Leftrightarrow$ $D\mathbf{c}(\mathbf{q})$ has full rank (along the solution trajectory): Force constraint independent.

△

Example 3.8.28 (MATLAB-integrators for pendelum equation in descriptive form).

MATLAB `ode15s`/`ode23t` applied on (3.8.20):

```
???  Error using ==> funfun/private/daeic12 at 77
This DAE appears to be of index greater than 1.

Error in ==> ode15s at 394
[y,yp,f0,dfdy,nFE,nPD,Jfac] = daeic12(odeFcn,odeArgs,...)
```

Idea: "‘Öutfox MATLAB”, test constraint (3.8.22)

$$\mathbf{M}\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}): \quad \mathbf{M} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{f}(\mathbf{y}) = \begin{pmatrix} y_3 \\ y_4 \\ -y_5 y_1 \\ -y_5 y_2 - g \\ -y_5(y_1^2 + y_2^2) - g y_2 + (y_3^2 + y_4^2) \end{pmatrix}.$$

- $l = 1$, $g = 9.8$, period of time $[0, 50]$, Solver: `ode15s` with default-tolerances
- consistent initial values $x_1(0) = -x_2(0) = \frac{1}{2}\sqrt{2}$, $p_1(0) = p_2(0) = 0$ ($\rightarrow \lambda(0)$)

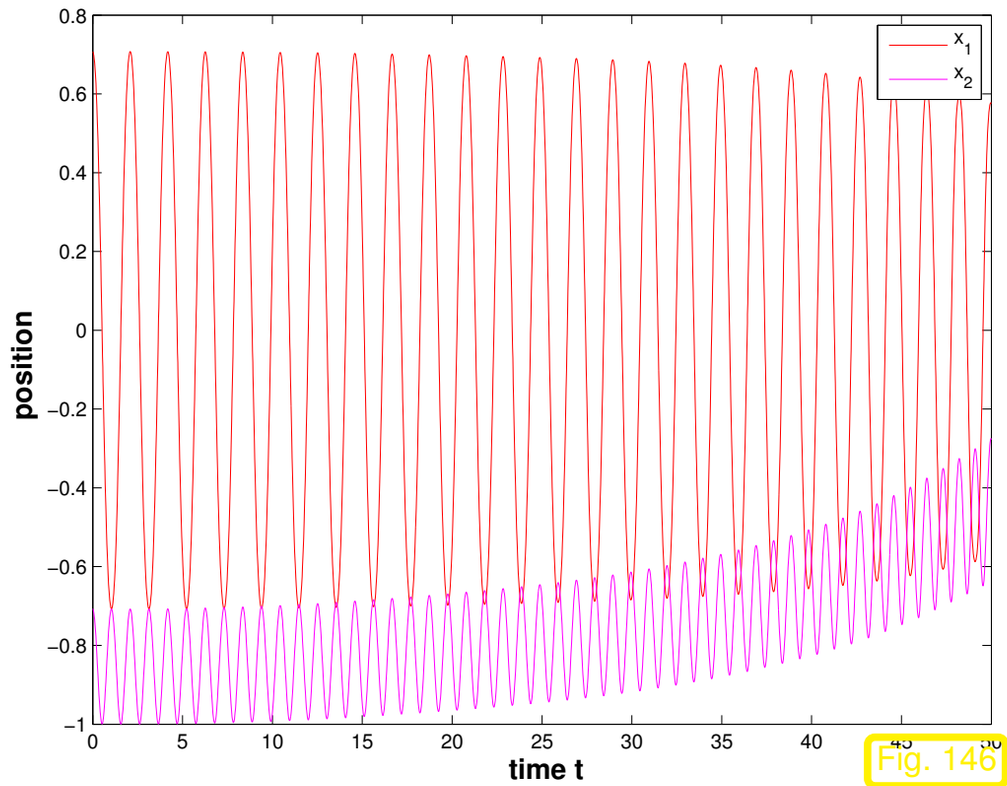


Fig. 146

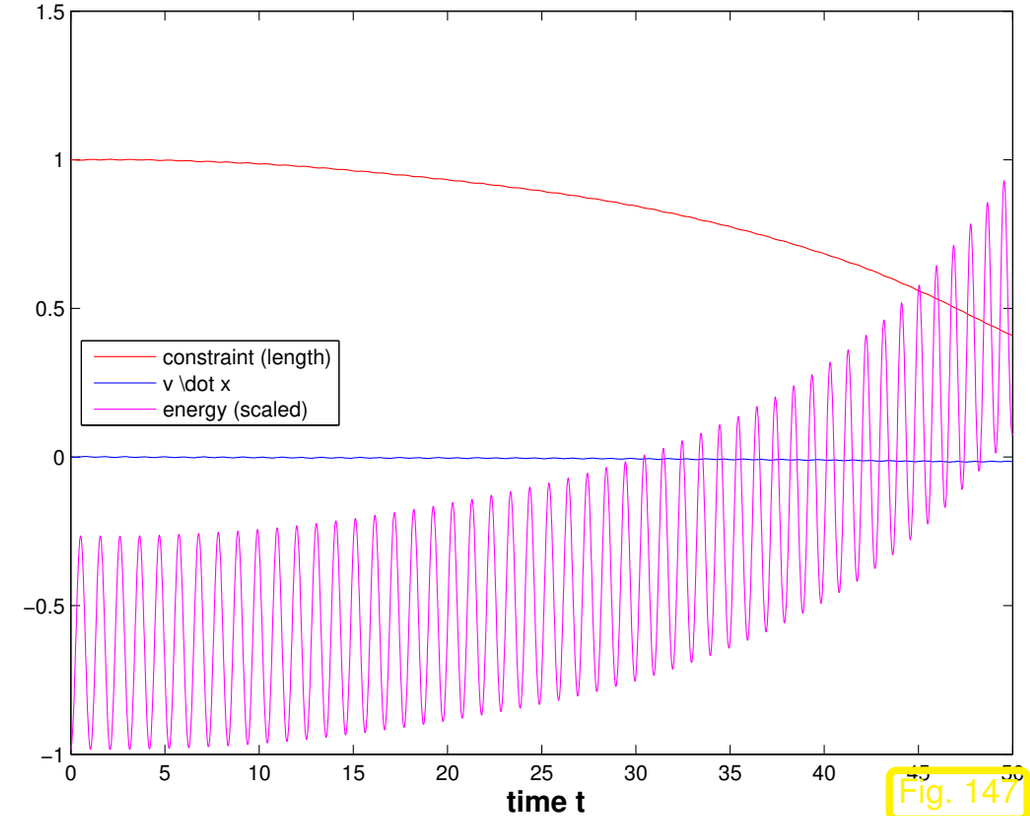


Fig. 147

◇

Example 3.8.29 (Implicit Euler-method for pendulum equations in descriptive form).

- IVP for pendulum-DAE (3.8.20) as in Ex. 3.8.28, endtime point $T = 5$
- Implicit Euler-method (1-level Radau-OSM)

Formal application of a backward differentiation quotient (→ Rem. 1.4.14) on the hamiltonian equation of motion with constraint force (3.8.24): compute $(\mathbf{q}_1, \mathbf{p}_1, \lambda_1)$ aus $(\mathbf{q}_0, \mathbf{p}_0, \lambda_0)$ according

$$\begin{aligned}
\mathbf{q}_1 &= \mathbf{q}_0 + h\frac{\partial H}{\partial \mathbf{p}}(\mathbf{p}_1, \mathbf{q}_1) \\
\mathbf{p}_1 &= \mathbf{p}_0 - h\frac{\partial H}{\partial \mathbf{q}}(\mathbf{p}_1, \mathbf{q}_1) - hD\mathbf{c}(\mathbf{q}_1)^T\boldsymbol{\lambda}_1 \\
0 &= \mathbf{c}(\mathbf{q}_1) \ .
\end{aligned}$$

concrete for pendulum equation in descriptive form (3.8.20), $\mathbf{q} \leftrightarrow \mathbf{x}$, $H(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\|\mathbf{p}\|^2 + gx_2$:

$$\begin{aligned}
\mathbf{x}_1 &= \mathbf{x}_0 + h\mathbf{p}_1 \ , \\
\mathbf{p}_1 &= \mathbf{p}_0 - h\left(\lambda\mathbf{x}_1 + \begin{pmatrix} 0 \\ g \end{pmatrix}\right) \ , \\
0 &= \|\mathbf{x}_1\|_2^2 - l^2 \ .
\end{aligned}$$

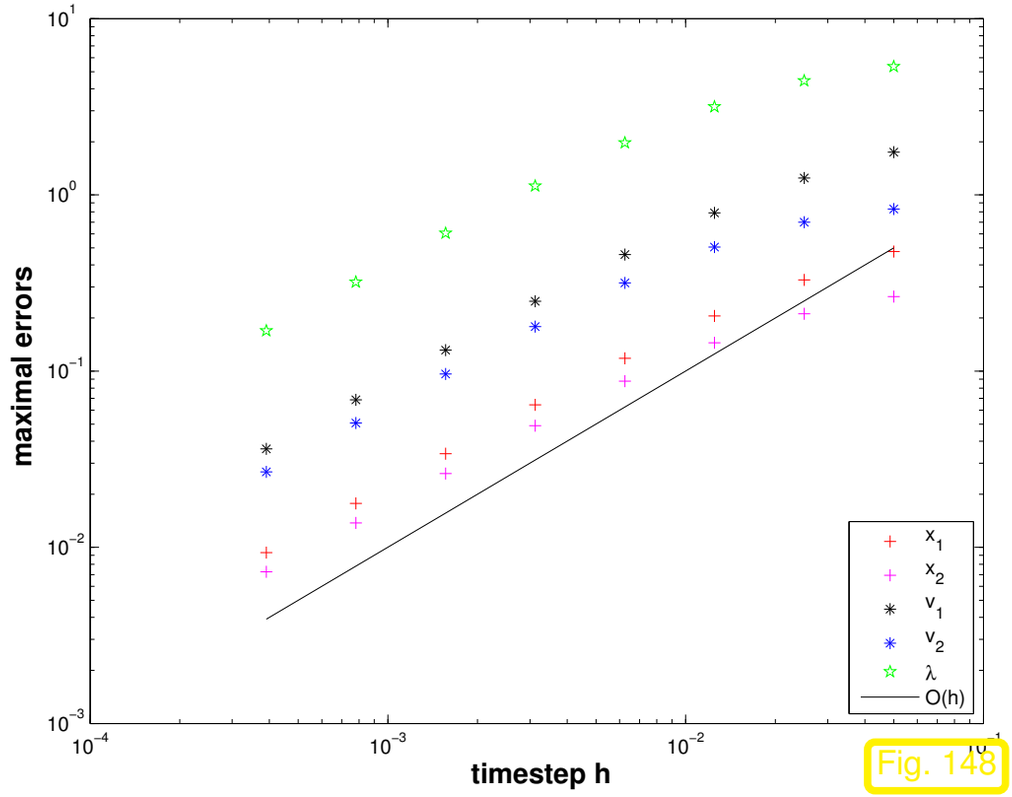


Fig. 148

◁ Errors in solution component (discrete maximum norm) for $100, 200, 400, 1600, 3200, 6400, 12800$ implicit Euler-steps

("'Exact solution computed with implicit midpoint rule applied on minimal coordinate form, see Ex. 1.4.24)

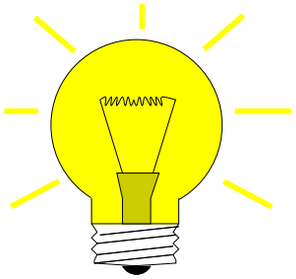
We observe algebraic *convergence of first order* in all solution components **!**

◇

P. Grohs

rev 63606, February 17, 2014

Sect. 3.8.2: Singular perturbation technique for Runge-Kutte-one-step method



Application on autonomous DAE (Index > 1!)

$$\begin{aligned} \dot{\mathbf{y}} &= \mathbf{f}(\mathbf{y}, \boldsymbol{\lambda}) \ , \\ 0 &= \mathbf{c}(\mathbf{y}) \ . \end{aligned} \tag{3.8.30}$$

($\mathbf{f} : D \times \mathbb{R}^q \mapsto \mathbb{R}^d$, $D \subset \mathbb{R}^d$, $\mathbf{c} : D \mapsto \mathbb{R}^q$, Assumption: consistent initial values $\mathbf{y}_0, \lambda_0$ at time $t = 0$)

To (3.8.30) belonging singular perturbed problem

$$\begin{aligned} \dot{\mathbf{y}} &= \mathbf{f}(\mathbf{y}, \boldsymbol{\lambda}) \ , \\ \epsilon \dot{\boldsymbol{\lambda}} &= \mathbf{c}(\mathbf{y}) \ , \end{aligned}$$

for $\epsilon \to 0$.

P. Grohs

rev 63606, February 17, 2014

Sect. 3.8.2 ➢ Consider stiffly-accurate RK-OSM

Formal computation: singular perturbation technique for Runge-Kutta-one-step-method with Butcher-scheme $\begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^T \end{array}$, $b_i = a_{s,i}$, $i = 1, \dots, s$:

$$\left\{ \begin{array}{l} \dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, \boldsymbol{\lambda}) \\ \dot{\boldsymbol{\lambda}} = \frac{1}{\epsilon}\mathbf{c}(\mathbf{y}) \end{array} \right. \quad \blacktriangleright \quad \left\{ \begin{array}{l} \left\{ \begin{array}{l} \mathbf{g}_i = \mathbf{y}_0 + h \sum_{j=1}^{s} a_{ij}\mathbf{f}(\mathbf{g}_j, \mathbf{g}_j^\lambda) \\ \epsilon\mathbf{g}_i^\lambda = \epsilon\boldsymbol{\lambda}_0 + h \sum_{j=1}^{s} a_{ij}\mathbf{c}(\mathbf{g}_j) \end{array} \right. , i = 1, \dots, s \\ \left\{ \begin{array}{l} \mathbf{y}_1 := \mathbf{y}_0 + h \sum_{i=1}^{s} a_{s,i}\mathbf{f}(\mathbf{g}_i, \mathbf{g}_i^\lambda) = \mathbf{g}_s \\ \mathbf{v}_1 := \mathbf{v}_0 + \frac{1}{\epsilon}h \sum_{i=1}^{s} a_{s,i}\mathbf{c}(\mathbf{g}_i) = \mathbf{g}_s^\lambda \end{array} \right. \end{array} \right.$$

$$\epsilon \to 0 \quad \Rightarrow \quad \sum_{j=1}^{s} a_{ij}\mathbf{c}(\mathbf{g}_j) = 0, \; i = 1, \dots, s \quad \Rightarrow \quad \mathbf{c}(\mathbf{g}_j) = 0$$

RK-OSM stiffly-accurate $\Rightarrow$ coefficient matrix $\mathfrak{A} := (a_{ij})_{i,j=1}^{s}$ regular

P. Grohs

rev 63606, February 17, 2014

➢ level equations (→ Rem. 2.3.7) for stiffly-accurate RK-OSM for (3.8.30)

$$\begin{cases} \mathbf{g}_i &= \mathbf{y}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{f}(\mathbf{g}_j, \mathbf{g}_j^{\lambda}) \\ 0 &= \mathbf{c}(\mathbf{g}_i) \end{cases} \quad , i = 1,\dots,s \; , \quad \mathbf{y}_1 = \mathbf{g}_s \; . \tag{3.8.31}$$

Note:
- (3.8.31) $\hat{=}$ implicit equation for level $\mathbf{g}_i, \mathbf{g}_i^{\lambda}$ ($s(d+q)$ Unknown)
- $\boldsymbol{\lambda}_0$ is not used!

Remark 3.8.32 (Implementing stiffly-accurate RK-OSM for DAE).

Formal: Incremental equation of a RK-OSM for *general* autonomous DAE $\mathbf{M}\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\mathbf{M} \in \mathbb{R}^{d,d}$ singular

$$\mathbf{M}\mathbf{k}_i = \mathbf{f}\big(\mathbf{y}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j\big) \; , \quad i = 1,\dots,s \; .$$

$$\Updownarrow$$

$$F(\mathfrak{k}) = 0 \quad , \quad F(\mathfrak{k}) = \begin{pmatrix} \mathbf{M}\mathbf{k}_1 - \mathbf{f}(\mathbf{y}_0 + h\sum_{j=1}^{s} a_{1j}\mathbf{k}_j) \\ \vdots \\ \mathbf{M}\mathbf{k}_s - \mathbf{f}(\mathbf{y}_0 + h\sum_{j=1}^{s} a_{sj}\mathbf{k}_j) \end{pmatrix} \; , \quad \mathfrak{k} = \begin{pmatrix} \mathbf{k}_1 \\ \vdots \\ \mathbf{k}_s \end{pmatrix} \; . \tag{3.8.33}$$

RK-OSM stiffly-accurate ▶ $\mathbf{y}_1 = \mathbf{y}_0 + h \sum_{j=1}^{s} a_{sj} \mathbf{k}_j$.

MATLAB-CODE : steif-genaues RK-ESV für DAE

```
function y1 = rksadaestep(rhs,M,y0,h,A)
  d = length(y0);
  s = size(A,1);
  F = @(gv) stagefn(gv,y,h,fun,A,M);
  [dgv,r,flg] = fsolve(F,zeros(s*d,1));
  if (flg <= 0), error('fsolve'); end
  y1 = dgv((s-1)*d+1:s*d);
end
```

◁ Single step: stiffly-accurate RK-OSM for $\mathbf{M}\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$

Coefficient matrix $\mathfrak{A} \in \mathbb{R}^{s,s}$

`fsolve` from MATLAB "optimization toolbox"

Extract s. Level $\mathbf{g}_s$

MATLAB-CODE : Stufen für steif-genaues RK-ESV

```
function dgv1 = stagefn(dgv0,y,h,fun,A,M);
  s = size(A,1); d = length(y);
  GV = reshape(dgv0,d,s);
  dgv1 = kron(eye(s),M)*dgv0;
  for j=1:s
    fg = feval(fun,y+GV(:,j));
    dgv1 = dgv1-h*kron(A(:,j),eye(d))*fg;
  end
```

◁ Evaluation of $F(\mathfrak{g} - \mathfrak{y}_0)$, see (3.8.33)

Convergence theory for (3.8.31) in the case of DAEs *with index 2*: [23, Sect. VII.4]

For s-level Radau-OSM: $\mathbf{y}_h(t)$ converges with order $2s-1$, $\boldsymbol{\lambda}_h(t)$ with order s.

3.8.4 Optimal Control Problems

4 Structure-preserving numerical Integration

Structure = essential *qualitative* properties of an evolution (☞ Sect. 1.3.3.5)

- first integrals/invariants ($\rightarrow$ Def. 1.2.7), e.g. total energy, torsional moment, see Sect. 1.2.4
- attractive and repelling fixpoints, see 3.2
- nonexpansivity, see 3.3
- new: special correspondece between flow and autonomous DE.
  ☞ conservation of volume, symplecticity, etc.

"Inheritance" of structure to discrete evolution Ψ

Important for long time integration of a "qualitatively correct solution"

P. Grohs

rev 63606, February 17, 2014

backward analysis

Numerical Solution is exact solution to a “structural similar problem” with slightly disturbed initial data/parameters Note: In this chapter we restrict ourselves to autonomous differential equations.

4.1 Polynomial Invariants

Reminder to Def. 1.2.7 & (1.2.8): Concept and properties of invariants/first integrals

Example: conservation of mass $\rightarrow$ Sect. 1.2.2, conservation of energy $\rightarrow$ Lemma 1.2.23, conservation of length Ex. 1.4.18

IVP for autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ on the state space $D \subset \mathbb{R}^d$

$t \mapsto \mathbf{y}(t) \hat{=}$ solution of initital value problem $\mathbf{y}_0 \in D$ first Integral $I : D \mapsto \mathbb{R}$ satisfies $I(\mathbf{y}(t)) = \mathrm{const}$ for *every* solution $\mathbf{y}(t)$ ($\rightarrow$ Def. 1.2.7)

(1.2.8): I is first integral of $\dot{\mathbf{y}} = f(\mathbf{y}) \quad \Leftrightarrow \quad \mathbf{grad}\, I(\mathbf{y}) \cdot \mathbf{f}(\mathbf{y}) = 0$ for all $\mathbf{y} \in D$.

linear first integral : $I(\mathbf{y}) = \mathbf{b}^T\mathbf{y} + c$ with $\mathbf{b} \in \mathbb{R}^d$, $c \in \mathbb{R}$
quadratic first integral : $I(\mathbf{y}) = \frac{1}{2}\mathbf{y}^T\mathbf{M}\mathbf{y} + \mathbf{b}^T\mathbf{y} + c$ mit $\mathbf{M} \in \mathbb{R}^{d,d}$, $\mathbf{b} \in \mathbb{R}^d$, $c \in \mathbb{R}$

Definition 4.1.1 (Polynomial Invariants)**.**
First Integral $I(\mathbf{y})$ is polynomial of degree n, $n \in \mathbb{N}$, if

$$I(\mathbf{y}) = \sum_{\boldsymbol{\alpha}\in\mathbb{N}_0^d,\, |\boldsymbol{\alpha}|\leq n} \beta_{\boldsymbol{\alpha}}\mathbf{y}^{\boldsymbol{\alpha}}\,, \quad \beta_{\boldsymbol{\alpha}} \in \mathbb{R} \quad \textit{(multivariate polynomial)}\,.$$

✎ Multiindexnotation: $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$, $|\boldsymbol{\alpha}| = \sum_i \alpha_i$, $\mathbf{y}^{\boldsymbol{\alpha}} := y_1^{\alpha_1} \cdot \dots \cdot y_d^{\alpha_d}$

Theorem 4.1.2 (Conservation of linear invariants)**.**
All Runge-Kutta One-Step methods ($\rightarrow$ Def. 2.3.5) conserve linear first integrals.

Proof. (for the autonomous case)

Linear invariant $I(\mathbf{y}) = \mathbf{b}^T\mathbf{y} + c$, $\mathbf{b} \in \mathbb{R}^d$, $c \in \mathbb{R}$ ➢ $\mathbf{grad}\, I(\mathbf{y}) = \mathbf{b} \quad \forall \mathbf{y} \in D$

$$\text{(1.2.8)} \quad \Rightarrow \quad \mathbf{b} \cdot \mathbf{f}(\mathbf{y}) = 0 \,,$$

$$\Rightarrow \quad \mathbf{b} \cdot \mathbf{f}(\mathbf{y}_0 + h \sum_{j=1}^{s} a_{ij}\mathbf{k}_j) = \mathbf{b} \cdot \mathbf{k}_i = 0 \,, \quad i = 1, \dots, s \quad \text{(for Inkremente)} \,,$$

$$\Rightarrow \quad I(\mathbf{y}_1) = \mathbf{b} \cdot \mathbf{y}_1 + c = \mathbf{b} \cdot \big(\mathbf{y}_0 + \sum_{i=1}^{s} b_i\mathbf{k}_i\big) + c = \mathbf{b} \cdot \mathbf{y}_0 + c = I(\mathbf{y}_0) \,. \qquad \square$$

Example 4.1.3 (Precision of a magnetic needle).

$\mathbf{y} : \mathbb{R} \mapsto \mathbb{R}^3$ = Velocity at the top of a magnetic needle (in the outer field $\mathbf{h}$, fixed at 0)

➢ equation of motion

$$\dot{\mathbf{y}} = \mathbf{y} \times \mathbf{h} \quad , \quad \text{Cross product} \quad \mathbf{y} \times \mathbf{h} = \begin{pmatrix} y_2h_3 - y_3h_2 \\ y_3h_1 - y_1h_3 \\ y_1h_2 - y_2h_1 \end{pmatrix} \perp \mathbf{y}$$

quadratic invariants: $\left\| \mathbf{y}^{(m)}(t) \right\| = \text{const} \;, \quad m \in \mathbb{N}_0$

Initial value: $\mathbf{y}_0 = (\frac{1}{2}\sqrt{2}, 0, 1, \frac{1}{2}\sqrt{2})^T$

P. Grohs

rev 63606, February 17, 2014

MATLAB-CODE : Computation of the precision of a magnetic needle

```
h = [-1;-1;-1]; tspan = [0 10000]; y0 = [0.5*sqrt(2);0;0.5*sqrt(2)];
fun = @(t,x) cross(x,h);
Jac = @(t,x) [0 h(3) -h(2); -h(3) 0 h(1); h(2) -h(1) 0];
options = odeset('reltol',0.001,'abstol',1e-4,'stats','on');
[t45,y45] = ode45(fun,tspan,y0,options);
options = odeset('reltol',0.001,'abstol',1e-4,'stats','on','Jacobian',Jac);
[t23,y23] = ode23s(fun,tspan,y0,options);
```

`ode45`: 24537 successful steps, 7432 failed attempts, 191815 function evaluations

`ode23s`: 93447 successful steps, 4632 failed attempts, 289607 function evaluations

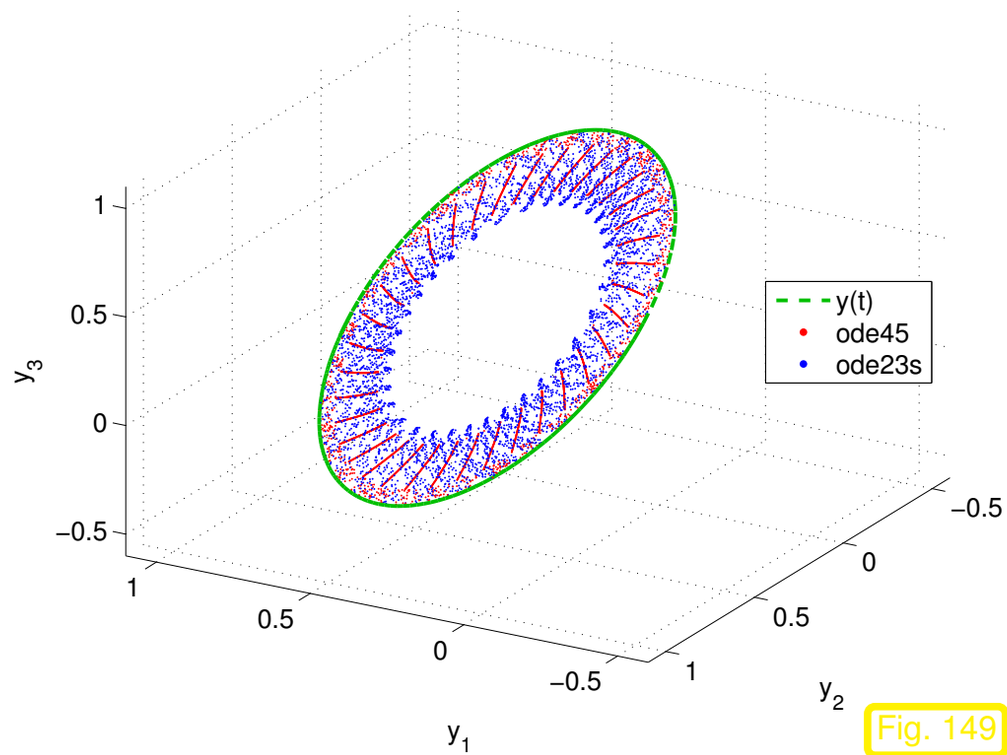


Fig. 149

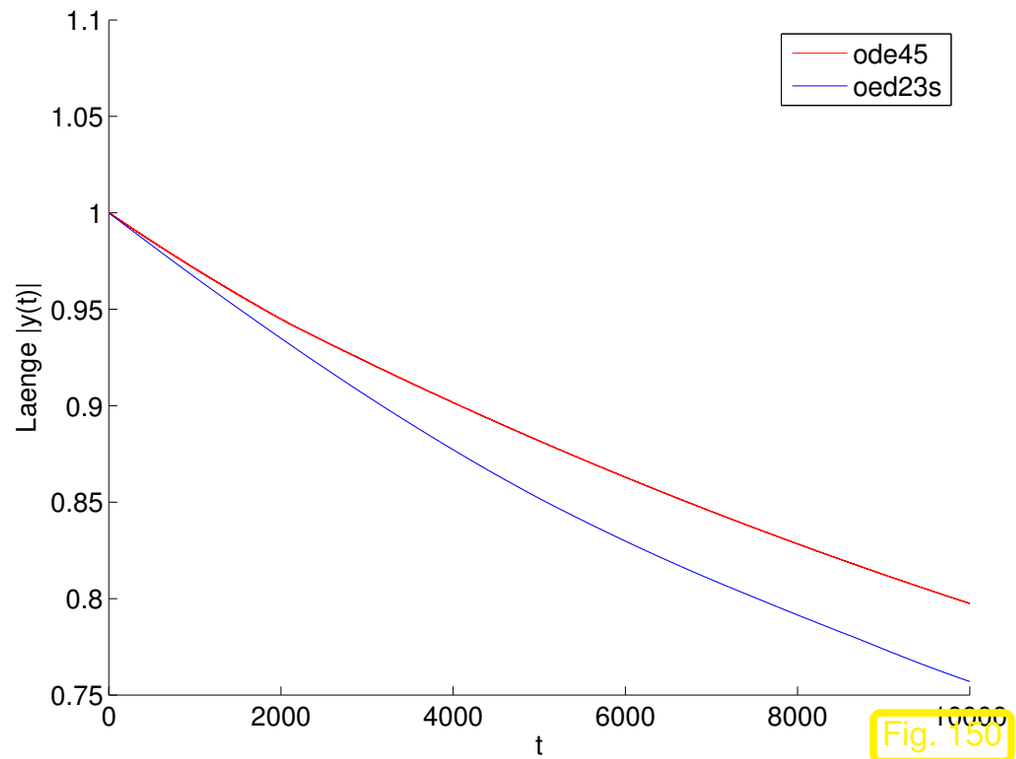


Fig. 150

No conservation of $\|\mathbf{y}(t)\|$ for a *long time* ($\leftrightarrow$ many periods of oscillation of the pendulum)

One-step method on an equidistant timemesh (qualitative behavior):

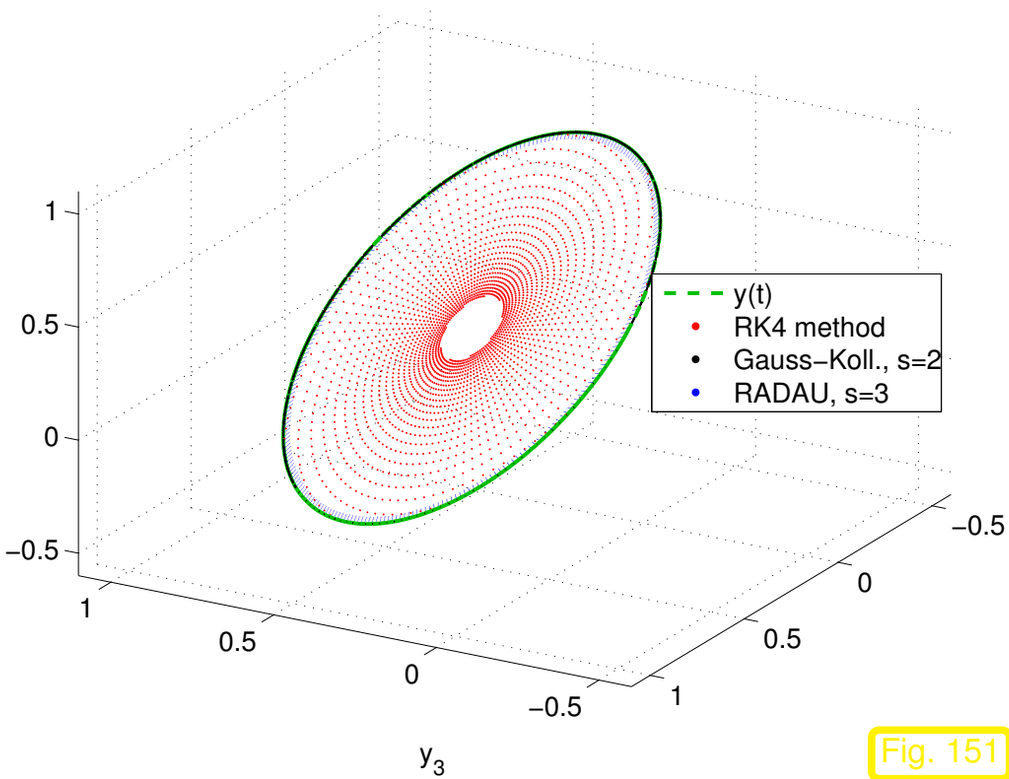


Fig. 151

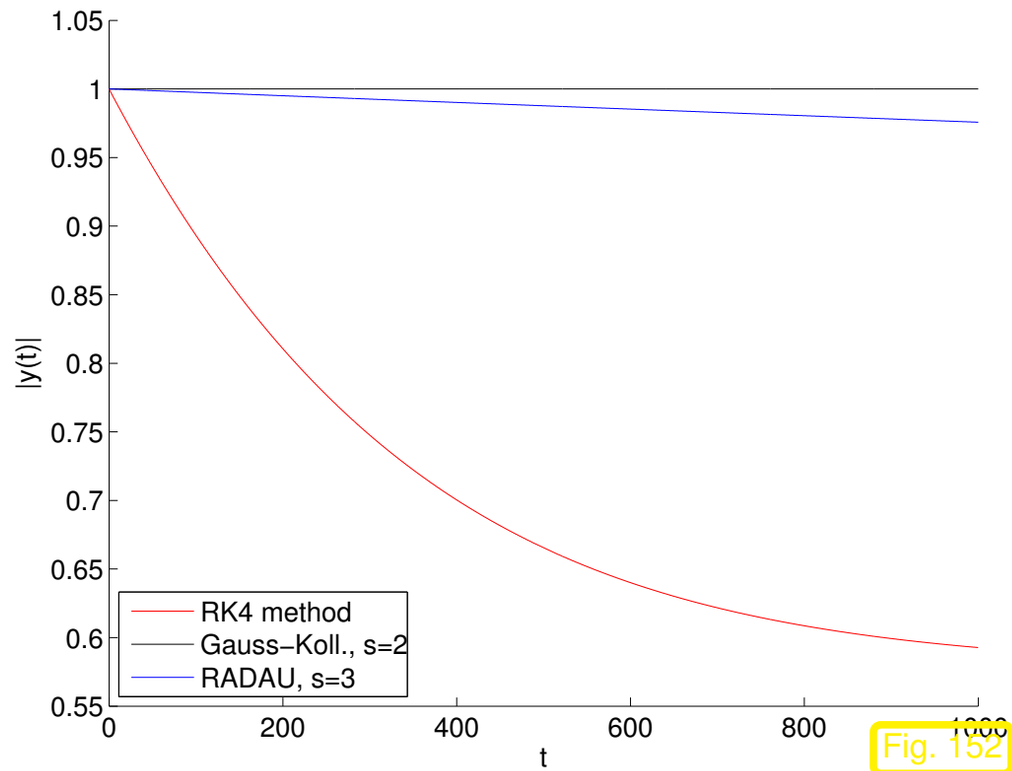


Fig. 152

P. Grohs

rev 63606, February 17, 2014

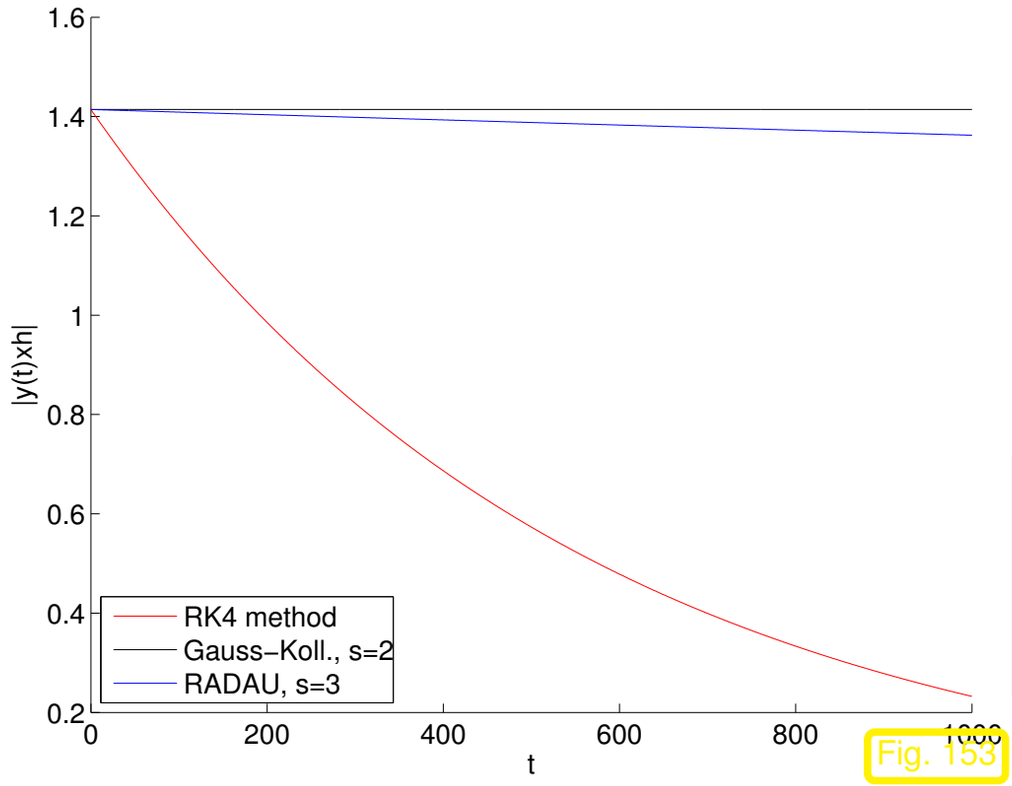


Fig. 153

△

Conservation of *all* quadratic invariants only by the Gauss-collocation method.

Reminder to Lemma 1.4.23: Conservation of quadratic first integral by the implicit midpoint rule, the simplest Gauss-collocation method

◇

P. Grohs

rev 63606, February 17, 2014

Theorem 4.1.4 (Conservation of quadratic invariants)**.**

Gauss-collocation-OSM (→ Sect. 2.2.3) conserves quadratic first integrals.

Proof. (for autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, cf. Proof of Thm. 3.3.7)

$\mathbf{y}_h(t) \in \mathcal{P}_s \hat{=}$ Gauss-collocation-polynomial to initial value $\mathbf{y}_0$ ($t_0 = 0$):

$$\dot{\mathbf{y}}_h(c_i h) = \mathbf{f}(\mathbf{y}_h(c_i h)) \ , \quad h \hat{=} \text{ stepwidth, cf. (2.2.1)} \ . \tag{4.1.5}$$

quadratic invariant: $I(\mathbf{y}) = \frac{1}{2}\mathbf{y}^T\mathbf{M}\mathbf{y} + \mathbf{b}^T\mathbf{y} + c$ with $\mathbf{M} = \mathbf{M}^T \in \mathbb{R}^{d,d}$, $\mathbf{b} \in \mathbb{R}^d$, $c \in \mathbb{R}$

$d(\tau) := I(\mathbf{y}_h(\tau h))$ is polynomial of degree $\leq 2s$.

Since s-point gauss-quadrature exact for polynomials of degree $\leq 2s-1$ and $d' \in \mathcal{P}_{2s-1}$

$$d(1) = d(0) + \int_0^1 d'(\tau)\,\mathrm{d}\tau = d(0) + \underbrace{\sum_{i=1}^{s} b_i d'(c_i)}_{\text{Goal } \overset{!}{=} 0} \ .$$

From collocation requirements (4.1.5) and (1.2.8)

$$d'(\tau) = h\,\mathbf{grad}\, I(\mathbf{y}_h(\tau h)) \cdot \dot{\mathbf{y}}_h(\tau h) \quad \Rightarrow \quad d'(c_i) = h\,\underbrace{\mathbf{grad}\, I(\mathbf{y}_h(c_i h)) \cdot \mathbf{f}(\mathbf{y}_h(c_i h))}_{=0} = 0 \ .$$

The claim follows from $d(0) = I(\mathbf{y}_0)$, $d(1) = I(\mathbf{y}_1)$. □

Lemma 4.1.6 (Conservation of quadratic invariants by RK-ESV)**.**
If the coefficients of a s-level (consistent) Runge-Kutta-one-step method ($\rightarrow$ Def. 2.3.5) satisfy

$$b_i a_{ij} + b_j a_{ji} = b_i b_j \quad \text{for all } i,j = 1,\dots,s \;, \tag{4.1.7}$$

then the discrete evolution quadratic first integrals.

Proof: (for simplified quadratic invariants $I(\mathbf{y}) := \frac{1}{2}\mathbf{y}^T\mathbf{M}\mathbf{y}$, $\mathbf{M} \in \mathbb{R}^{d,d}$, $\mathbf{M} = \mathbf{M}^T$)

$$\text{(1.2.8)} \quad \succ \qquad \mathbf{grad}\, I(\mathbf{y}) = \mathbf{M}\mathbf{y} \quad \Rightarrow \quad \mathbf{y}^T\mathbf{M}\mathbf{f}(\mathbf{y}) = 0 \quad \forall \mathbf{y} \in D \;. \tag{4.1.8}$$

One step of a RK-ESV with increments $\mathbf{k}_i$, cf. Def. 2.3.5:

$$\mathbf{y}_1 = \mathbf{y}_0 + h\sum_{i=1}^{s} b_i \mathbf{k}_i \;,$$

$$\blacktriangleright \quad \mathbf{y}_1^T\mathbf{M}\mathbf{y}_1 - \mathbf{y}_0^T\mathbf{M}\mathbf{y}_0 = 2h\sum_{i=1}^{s} b_i \mathbf{y}_0^T\mathbf{M}\mathbf{k}_i + h^2\sum_{i=1}^{s}\sum_{j=1}^{s} b_i b_j \mathbf{k}_i^T\mathbf{M}\mathbf{k}_j \;. \tag{4.1.9}$$

Use Levelform, cf. Rem. 2.2.5:

$$\mathbf{g}_i = \mathbf{y}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{f}(\mathbf{g}_j) \quad \succ \quad \mathbf{k}_i = \mathbf{f}(\mathbf{g}_i),\, i=1,\dots,s, \qquad \boxed{\mathbf{y}_0 = \mathbf{g}_i - h\sum_{j=1}^{s} a_{ij}\mathbf{f}(\mathbf{g}_j).}$$

Plugging in (4.1.9), since by (4.1.8) it follows that $\mathbf{g}_i\mathbf{M}\mathbf{f}(\mathbf{g}_i) = 0$:

$$\begin{aligned}
\mathbf{y}_1^T\mathbf{M}\mathbf{y}_1 - \mathbf{y}_0^T\mathbf{M}\mathbf{y}_0 &= 2h\sum_{i=1}^{s} b_i\left(\mathbf{g}_i - h\sum_{j=1}^{s} a_{ij}\mathbf{f}(\mathbf{g}_j)\right)^T \mathbf{M}\mathbf{f}(\mathbf{g}_i) + h^2\sum_{i=1}^{s}\sum_{j=1}^{s} b_i b_j \mathbf{f}(\mathbf{g}_i)^T\mathbf{M}\mathbf{f}(\mathbf{g}_j) \\
&= -2h^2\sum_{i=1}^{s} b_i\sum_{j=1}^{s} a_{ij}\mathbf{f}(\mathbf{g}_j)^T\mathbf{M}\mathbf{f}(\mathbf{g}_i) + h^2\sum_{i=1}^{s}\sum_{j=1}^{s} b_i b_j \mathbf{f}(\mathbf{g}_i)^T\mathbf{M}\mathbf{f}(\mathbf{g}_j) \\
&= h^2\sum_{i=1}^{s}\sum_{j=1}^{s}(-2b_i a_{ij} + b_i b_j)\mathbf{f}(\mathbf{g}_i)^T\mathbf{M}\mathbf{f}(\mathbf{g}_j)\,.
\end{aligned}$$

$\mathbf{M} = \mathbf{M}^T$ $\succ$ index swap in the doublesum $\succ$ Claim. $\square$

Theorem 4.1.10 (Non-conservation of general polynomial invariants)**.**
For $n \geq 3$ there is no *consistent Runge-Kutta one-step method ($\rightarrow$ Def. 2.3.5) which for all autonomous differential equation $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ conserves all its polynomial invariantens ($\rightarrow$ Def. 4.1.1) of degree n.*

Lemma for the proof:

Lemma 4.1.11 (Derivation of the determinant function)**.**
For the determinant-function $\det : \mathbb{R}^{d,d} \mapsto \mathbb{R}$ we have

$$(D\det(\mathbf{X}))(\mathbf{H}) = \operatorname{trace}(\operatorname{adj}(\mathbf{X})\mathbf{H}) \;, \quad \mathbf{X}, \mathbf{H} \in \mathbb{R}^{d,d} \;.$$

P. Grohs

rev 63606, February 17, 2014

Trace of a matrix $\mathbf{A} = (a_{ij})_{i,j=1}^{d} \in \mathbb{R}^{d,d}$: $\operatorname{trace}(\mathbf{A}) = \sum_{j=1}^{d} a_{jj}$

adjoint matrix $(\operatorname{adj}(\mathbf{X}))_{ij} = (-1)^{i+j}\det(\check{\mathbf{X}}_{ij})$, $\mathbf{X} \in \mathbb{R}^{d,d}$, $1 \leq i,j \leq d$, $\check{\mathbf{X}}_{ij} \hat{=}$ matrix, which we got from $\mathbf{X}$ by removing i. row and j. column (minor).

☞ Known from linear algebra [14, Lemma 4.3.4]: $\mathbf{A} \cdot \operatorname{adj}(\mathbf{A}) = \det(\mathbf{A}) \cdot \mathbf{I}$

Proof. As a polynomial in the matrix-elements the map $\mathbf{A} \mapsto \det \mathbf{A}$ is a C^∞-function:

$$\det \mathbf{A} := \sum_{\sigma \in \Pi_d} \operatorname{sgn}(\sigma) \prod_{i=1}^{d} a_{i,\sigma(i)} \,.$$

$$\blacktriangleright \quad \det(\mathbf{I} + \epsilon \mathbf{H}) = \sum_{\sigma \in \Pi_d} \operatorname{sgn}(\sigma) \prod_{i=1}^{d} (\delta_{i,\sigma(i)} + \epsilon h_{i,\sigma(i)})$$

$$= \prod_{i=1}^{d} (1 + \epsilon h_{ii}) + O(\epsilon^2) = 1 + \epsilon \sum_{i=1}^{d} h_{ii} + O(\epsilon^2) \,. \,,$$

for $\mathbf{H} = (h_{ij})_{i,j=1}^{d}$, since every permutation $\neq Id$ produces a product of size $O(\epsilon^2)$. Hence for regular $\mathbf{X} \in \mathbb{R}^{d,d}$:

$$\det(\mathbf{X} + \epsilon \mathbf{H}) - \det(\mathbf{X}) = \epsilon \operatorname{trace}(\underbrace{\det(\mathbf{X}) \mathbf{X}^{-1}}_{\operatorname{adj}(\mathbf{X})} \mathbf{H}) + O(\epsilon^2) \,.$$

Since the regular matrices in $\mathbb{R}^{d,d}$ lie dense, $\mathbf{X} \mapsto \operatorname{adj} \mathbf{X}$ is continuous $\rightarrow$ □

Proof of Thm. 4.1.10 (Proof by contradiction)

$t \mapsto \mathbf{Y}(t)$ solve linear Matrix-differential equation $\dot{\mathbf{Y}} = \mathbf{A}\mathbf{Y}, \quad \mathbf{A} \in \mathbb{R}^{d,d}$

▶ With Lemma 4.1.11 if follows for $I(\mathbf{Y}) = \det \mathbf{Y}$ that

$$\begin{aligned} D_{\mathbf{Y}} I(\mathbf{Y})\mathbf{H} &= \det \mathbf{Y} \cdot \operatorname{trace}(\mathbf{Y}^{-1}\mathbf{H}) \\ \Rightarrow \quad \frac{d}{dt} \det \mathbf{Y}(t) &= \det \mathbf{Y} \cdot \operatorname{trace}(\dot{\mathbf{Y}}\mathbf{Y}^{-1}) = \det \mathbf{Y} \cdot \operatorname{trace}(\mathbf{A}) \,. \end{aligned} \tag{4.1.12}$$

If $\operatorname{trace}(\mathbf{A}) = 0$ is $I(\mathbf{Y}) := \det \mathbf{Y}$ a polynomial invariant of degree d of the matrix-differential equation $\dot{\mathbf{Y}} = \mathbf{A}\mathbf{Y}$.

Assumption: RK-OSM contains polynomial invariants of degree $d > 2$. Apply the method on $\dot{\mathbf{Y}} = \mathbf{A}\mathbf{Y}$, $\operatorname{trace}(\mathbf{A}) = 0$, $\mathbf{A} \in \mathbb{R}^{d,d}$. By Rem. 3.1.13, (3.1.16)

$$\mathbf{Y}_1 = S(h\mathbf{A})\mathbf{Y}_0 \quad \text{mit stability function } S(z)\,, \quad h > 0\,.$$

▶ $$\det \mathbf{Y}_1 = \det \mathbf{Y}_0 \quad \forall \mathbf{Y}_0 \quad \Rightarrow \quad \det S(h\mathbf{A}) = 1$$

Choose special (diagonal !) matrix with $\operatorname{trace}(\mathbf{A})$ = 0 and time stepwidth $h = 1$

$$\mathbf{A} = \operatorname{diag}(\mu, \nu, -(\mu+\nu), 0, \dots, 0) \in \mathbb{R}^{d,d}\,, \quad \mu, \nu \in \mathbb{R}\,.$$

$$S(\mathbf{A}) = \operatorname{diag}(S(\mu), S(\nu), S(-(\mu+\nu)), 0, \dots, 0)$$

From $\det S(\mathbf{A}) = 1$ it follows that S satisfies the functional equation $S(\mu)S(\nu)S(-(\mu+\nu)) = 1$.

$$\Rightarrow \quad S(0) = 1 \quad \Rightarrow \quad S(-\mu) = S(\mu)^{-1} \quad \Rightarrow \quad S(\mu)S(\nu) = S(\mu+\nu) \quad \forall \mu, \nu \in \mathbb{R}\,.$$

$z \mapsto S(z)$ satisfies the functional equation of the exponential function, is continuous in an environment of $0 \quad \Rightarrow \quad S(z) = \exp(z)$.

On the other hand $S(z)$ has to be a rational function, see Thm. 3.1.6, a **contradiction** $\square$

4.2 Conservation of Volume

Physics: incompressible Flow $\leftrightarrow$ volume-preserving flow

Definition 4.2.1 (Conservation of volume)**.**
A map $\mathbf{\Phi} : D \subset \mathbb{R}^d \mapsto \mathbb{R}^d$ *is called* *volume-preserving*

$$\forall V \subset D \text{ measurable}: \quad \operatorname{Vol}(\mathbf{\Phi}(V)) = \operatorname{Vol}(V)\,.$$

Lemma 4.2.2 (volume-preserving map)**.**
A continuously differentiable map $\mathbf{\Phi} : D \subset \mathbb{R}^d \mapsto \mathbb{R}^d$ *is volume-preserving if and only if* $|\det D\mathbf{\Phi}(\mathbf{y})| = 1$ *for all* $\mathbf{y} \in D$*.*

Proof. We use an integral transformation:

$$\mathrm{Vol}(\mathbf{\Phi}(V)) = \int_{\mathbf{\Phi}(V)} 1 \,\mathrm{d}\mathbf{x} = \int_V |\det D\mathbf{\Phi}(\mathbf{y})| \,\mathrm{d}\mathbf{y} \,. \qquad \square$$

P. Grohs

rev 63606, February 17, 2014

Theorem 4.2.3 (Liouville's Theorem)**.**
Let $\mathbf{f} : D \subset \mathbb{R}^d \mapsto \mathbb{R}^d$ *be contionously differentiable. If and only if* $\mathrm{div}\,\mathbf{f}(\mathbf{y}) = 0$ *for all* $\mathbf{y} \in D$ *the evolution* $\mathbf{\Phi}^t$ *corresponding to* $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ *volume-preserving, i.e.*

$$\forall V \subset D \text{ compact}: \quad \exists \delta > 0: \quad \mathrm{Vol}(\mathbf{\Phi}^t(V)) = \mathrm{Vol}(V) \quad \forall 0 \le t < \delta \,.$$

✎ Notation: divergence $\mathrm{div}\,\mathbf{f}(\mathbf{y}) = \sum_{j=1}^{d} \frac{\partial f_i}{\partial y_i}(\mathbf{y}) = \mathrm{trace}\, D\mathbf{f}(\mathbf{y})$, with $\mathbf{f} = (f_1, \ldots, f_d)^T$

Proof. (based on lemma 4.2.2, cf. [21, Lemma 9.1])

Let $\mathbf{\Phi} : \widetilde{\Omega} \mapsto D$ be the evolution-operator to the autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$.

Jacobi-Matrix (propagation-matrix) $\mathbf{W}(t,\mathbf{y}) = D_{\mathbf{y}}\mathbf{\Phi}^t(\mathbf{y})$, $\mathbf{y} \in D$, satisfies variational equation (1.3.34)

$$\dot{\mathbf{W}}(t,\mathbf{y}) := \frac{d}{dt}\mathbf{W}(t,\mathbf{y}) = D\mathbf{f}(\mathbf{\Phi}^t\mathbf{y})\mathbf{W}(t,\mathbf{y}) \;, \quad t \in J(\mathbf{y}) \quad , \quad \mathbf{W}(0,\mathbf{y}) = \mathbf{I} \;. \tag{4.2.4}$$

As in the proof of Thm. 4.1.10, on Lemma 4.1.11, cf. (4.1.12):

$$\begin{aligned} \frac{d}{dt}\det \mathbf{W}(t,\mathbf{y}) &= \det \mathbf{W}(t,\mathbf{y}) \operatorname{trace}(\dot{\mathbf{W}}(t,\mathbf{y})\mathbf{W}^{-1}(t,\mathbf{y})) \\ &\overset{(??)}{=} \det \mathbf{W}(t,\mathbf{y}) \operatorname{trace}(D\mathbf{f}(\mathbf{\Phi}^t\mathbf{y})) \\ &= \det \mathbf{W}(t,\mathbf{y}) \operatorname{div} \mathbf{f}(\mathbf{\Phi}^t\mathbf{y}) \;. \end{aligned} \tag{4.2.5}$$

"$\Rightarrow$" aus (4.2.6), since $\det \mathbf{W}(0,\mathbf{y}) = 1$

"$\Leftarrow$": If $\operatorname{div}\mathbf{f} \neq 0$, there exists $\delta > 0$, $V \subset D$ such that $|\operatorname{div}\mathbf{f}(\mathbf{y})| > \delta$ for all $\mathbf{y} \in V$. Hence for $\mathbf{y} \in V$,

$$\frac{d}{dt}\det \mathbf{W}(t,\mathbf{y}) \geq \delta \det \mathbf{W}(t,\mathbf{y}) \quad \text{or} \quad \frac{d}{dt}\det \mathbf{W}(t,\mathbf{y}) \leq -\delta \det \mathbf{W}(t,\mathbf{y}) \;.$$

Lemma 1.3.29 $\Rightarrow$ $t \mapsto \det \mathbf{W}(t,\mathbf{y})$ increases/decreases exponentially. □

Incompressible Flow $\leftrightarrow$ divergence-free velocity-fields

Example 4.2.6 (Visualization of flow).

Application of numerical ODE-solver in *computergraphics*:

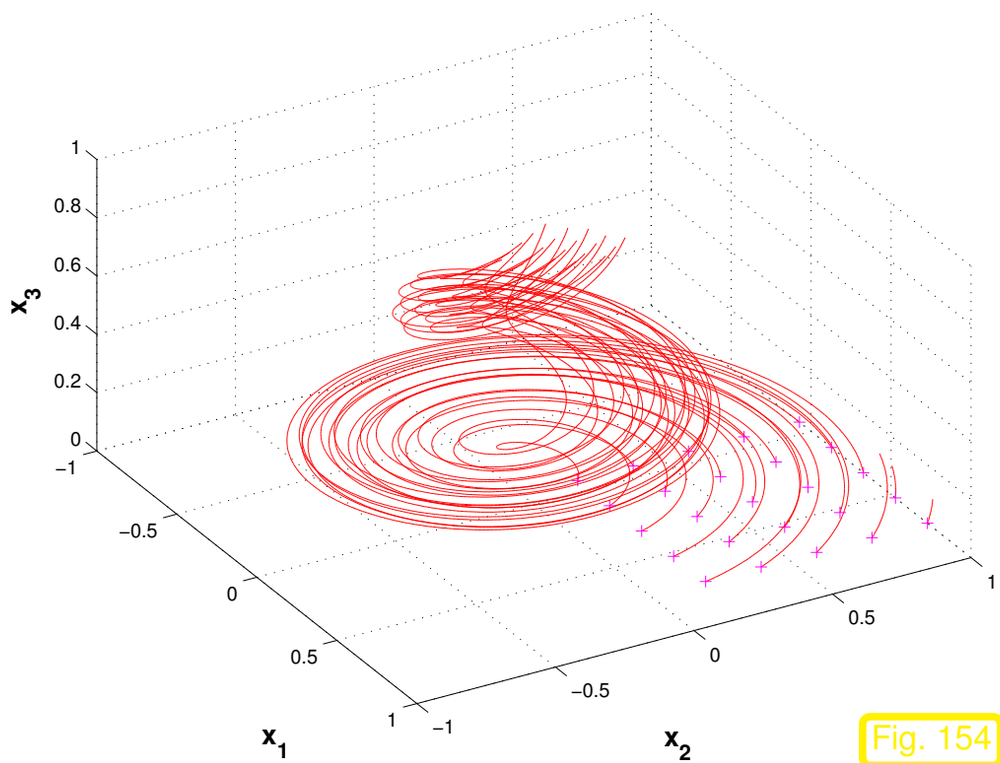


Fig. 154

divergencefree vectorfield:

$$\mathbf{f}(\mathbf{y}) = \begin{pmatrix} -y_2 - \frac{y_1}{a^2+y_3^2} \\ y_1 - \frac{y_2}{a^2+y_3^2} \\ 2/a \arctan(y_3/a) \end{pmatrix} , \quad \mathbf{y} \in \mathbb{R}^3 .$$

MATLAB-function

```
streamline(X,Y,Z,U,V,W,...)
```

Stromlinien von $\mathbf{f}$ $\hat{=}$ Solution of IVP for autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$.

Lemma 4.2.7 (variational equation und Runge-Kutta-one-step method)**.**
For Runge-Kutta-one-step method (→ Def. 2.3.5) the following diagram commutes

variational equation, see Sect. 1.3.3.4.

$$\begin{tikzcd}
\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}),\ \mathbf{y}(0) = \mathbf{y}_0 \arrow[r, "\mathbf{W} := \frac{\partial \mathbf{y}}{\partial \mathbf{y}_0}"] \arrow[d, "\textit{RK-ESV}"'] & \begin{array}{c} \dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}),\ \mathbf{y}(0) = \mathbf{y}_0\ , \\ \dot{\mathbf{W}} = D\mathbf{f}(\mathbf{y})\mathbf{W},\ \mathbf{W}(0) = \mathbf{I} \end{array} \arrow[d, "\textit{RK-ESV}"] \\
(\mathbf{y}_k)_{k=1}^N \arrow[r, "\mathbf{W}_k := \frac{\partial \mathbf{y}_k}{\partial \mathbf{y}_0}"] & (\mathbf{y}_k, \mathbf{W}_k)_{k=1}^N
\end{tikzcd}$$

Proof: (only for explicit euler-methods (1.4.2), cf. [21, Lemma 4.1])

Recursion of explicit euler-method for $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$

$$\mathbf{y}_{k+1} = \mathbf{y}_k + h\mathbf{f}(\mathbf{y}_k) \quad \overset{\frac{d}{d\mathbf{y}_0}}{\Longrightarrow} \quad \frac{d\mathbf{y}_{k+1}}{d\mathbf{y}_0} = \frac{d\mathbf{y}_k}{d\mathbf{y}_0} + hD\mathbf{f}(\mathbf{y}_k)\frac{d\mathbf{y}_k}{d\mathbf{y}_0}\ .$$

explicit euler-method for (extended) variational equation $\dot{\mathbf{W}} = D\mathbf{f}(\mathbf{y})\mathbf{W}$:

$$\mathbf{W}_{k+1} = \mathbf{W}_k + hD\mathbf{f}(\mathbf{y}_k)\mathbf{W}_k \ .$$

► $\frac{d\mathbf{y}_k}{d\mathbf{y}_0}$ and $\mathbf{W}_k$ satisfy the same recursion. □

The proof in the general case uses implicit differences of the runga-kutta incremental equation, see Def. 2.3.5.

Remark 4.2.8 (volume preserving integrators for $d = 2$).

For RK-ESV in the case $d = 2$:

$$\text{Conservation of quadratic invariants} \Longrightarrow \quad \text{Conservation of volume}$$

For $d = 2$: $\det \mathbf{W} = w_{11}w_{22} - w_{12}w_{21}$ is a quadratic function ($\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix} \hat{=}$ propagation matrix/Wronski-Matrix, see (1.3.33)). Is the evolution to $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ volume-preserving, we have

$\det \mathbf{W}(t) \equiv 1$, hence $\det \mathbf{W}$ is a quadratic invariant of the variational equation and is conserved by the RK-OSM.

By Lemma 4.2.8 we have

$$\forall h > 0: \quad \det \left(D_{\mathbf{y}_0} \mathbf{\Psi}^h \right) = 1 \ .$$

By Lemma 4.2.2 the discrete evolution is volume-preserving.

△

General Runge-Kutta-one-step methods are not considered **!** (Necessary: Integrators with "additional information" about $\mathbf{f}$)

Additive decomposition of $\mathbf{f}$ in *substantially two-dimensional* vector fields Splitting-method (→ Sect. 2.5 based on RK-ESV, the quadratic invariants conserved, see Sect. 4.1.

To ❶: $\mathbf{f}(\mathbf{y}) = \sum_{i=1}^{d-1} \mathbf{g}_{i,i+1}(\mathbf{y})$, $\mathbf{g}_{i,i+1}(\mathbf{y}) = (0 \ \cdots \ 0 \ \underset{\substack{\uparrow \\ i}}{*} \ \underset{\substack{\uparrow \\ i+1}}{*} \ 0 \ \cdots \ 0)^T$, $\operatorname{div} \mathbf{g}_{i,i+1} = 0$.

To ❷: $\mathbf{y}_{k+1} = (\mathbf{\Psi}_{d-1}^h \circ \cdots \circ \mathbf{\Psi}_1^h)\mathbf{y}_k$,

where $\mathbf{\Psi}_i^h \hat{=}$ discrete evolution of the RK-basismethod for $\dot{\mathbf{y}} = \mathbf{g}_{i,i+1}(\mathbf{y})$.

Generalized Lie-Trotter-Splitting (2.5.2)

existence of vectorfields $\mathbf{g}_{i,i+1}$ **?**

Theorem 4.2.9 (Decomposition of divergence-free vector-fields)**.**
Each continuous divergence-free $\mathbf{f} : \mathbb{R}^d \mapsto \mathbb{R}^d$ *can be written as a sum of* $d-1$ divergence-free *vector-fields* $\mathbf{g}_{i,i+1} : \mathbb{R}^d \mapsto \mathbb{R}^d$ *of the form*

$$\mathbf{g}_{i,i+1}(\mathbf{y}) = (0 \ \cdots \ 0 \ \underset{\substack{\uparrow \\ i}}{p_i(\mathbf{y})} \ \underset{\substack{\uparrow \\ i+1}}{q_i(\mathbf{y})} \ 0 \ \cdots \ 0)^T , \quad i = 1, \ldots, d-1 ,$$

with functions $p_i, q_i : \mathbb{R}^d \mapsto \mathbb{R}$.

P. Grohs

rev 63606, February 17, 2014

$$\mathbf{f}(\mathbf{y}) = \begin{pmatrix} f_1(\mathbf{y}) \\ f_2(\mathbf{y}) \\ f_3(\mathbf{y}) \\ f_4(\mathbf{y}) \\ \vdots \\ f_{d-1}(\mathbf{y}) \\ f_d(\mathbf{y}) \end{pmatrix} = \begin{pmatrix} p_1(\mathbf{y}) \\ q_1(\mathbf{y}) \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ p_2(\mathbf{y}) \\ q_2(\mathbf{y}) \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ p_3(\mathbf{y}) \\ q_3(\mathbf{y}) \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \cdots + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ p_{d-2}(\mathbf{y}) \\ q_{d-2}(\mathbf{y}) \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ p_{d-1}(\mathbf{y}) \\ q_{d-1}(\mathbf{y}) \end{pmatrix} .$$

Proof. (cf. [21, Theorem 9.3]) Constructive for $\mathbf{f} = (f_1, \dots, f_d)^T$ with arbitrary $a_i \in \mathbb{R}$

$$p_i(\mathbf{y}) = f_i(\mathbf{y}) + r_i(\mathbf{y}) \quad , \quad q_i(\mathbf{y}) = -r_{i+1}(\mathbf{y}) \ ,$$

$$r_i(\mathbf{y}) = \int_{a_i}^{y_i} \left(\frac{\partial f_1}{\partial y_1} + \cdots + \frac{\partial f_{i-1}}{\partial y_{i-1}} \right) (y_1, \dots, y_{i-1}, \tau, y_{i+1}, \dots, y_d) \, \mathrm{d}\tau \ , \quad 2 \le i \le d-1 \ ,$$

$$r_1(\mathbf{y}) \equiv 0 \ ,$$

$$\Rightarrow \quad \frac{\partial p_i}{\partial y_i} = \frac{\partial f_i}{\partial y_i} + \frac{\partial f_1}{\partial y_1} + \cdots + \frac{\partial f_{i-1}}{\partial y_{i-1}} \ , \qquad 1 \le i \le d-1 \ .$$

$$\frac{\partial q_i}{\partial y_{i+1}} = -\frac{\partial f_1}{\partial y_1} - \cdots - \frac{\partial f_i}{\partial y_i} = -\frac{\partial p_i}{\partial y_i} \ .$$

Hence all parts are divergence-free. Furthermore, since $\operatorname{div} \mathbf{f} = 0$,

$$\left(\sum_{i=1}^{d-1} \mathbf{g}_{i,i+1}(\mathbf{y})\right)_d = q_{d-1}(\mathbf{y}) = -r_d(\mathbf{y}) = -\int_{a_d}^{y_d} \frac{\partial f_1}{\partial y_1} + \cdots + \frac{\partial f_{d-1}}{\partial y_{d-1}} = \int_{a_d}^{y_d} \frac{\partial f_d}{\partial y_d} = f_d(\mathbf{y}) \ .$$

Note: construction of $\mathbf{g}_{i,i+1}$ requires (symbolic) Derivations of f_i.

Remark 4.2.10 (Volume-preserving Splitting-method of order 2).

Simplest volume-preserving Splitting-method of order 2:

- basis-RK-OSM: implicit mid-point-rule (1.4.19), $\mathbf{\Psi}_i^h \hat{=}$ discrete evolution to $\dot{\mathbf{y}} = \mathbf{g}_{i,i+1}(\mathbf{y})$
  ➥ $\mathbf{\Psi}_i^h$ volume-preserving, see Rem. 4.2.9 !
- Generalized Strang-Splitting (2.5.3)

$$\mathbf{\Psi}^h := \mathbf{\Psi}_1^{h/2} \circ \mathbf{\Psi}_2^{h/2} \circ \cdots \circ \mathbf{\Psi}_{d-2}^{h/2} \circ \mathbf{\Psi}_{d-1}^{h} \circ \mathbf{\Psi}_{d-2}^{h/2} \circ \cdots \circ \mathbf{\Psi}_1^{h/2} \ .$$

▶ symmetric One-step-method $\rightarrow$ Def. 2.1.27 $\overset{\text{Thm. 2.1.29}}{\Rightarrow}$ Order of consistency ≥ 2

△

4.3 Generalized Reversibility

reversible (symmetric) discrete evolutions/One-step methods $\longrightarrow$ Def. 2.1.27

(for autonomuous ODE) $\boxed{\boldsymbol{\Psi}^{-h} \circ \boldsymbol{\Psi}^{h} = Id \quad \text{for } h > 0 \text{ sufficiently small}}$

reversibility

$\Updownarrow$

Symmetry regarding time reversal

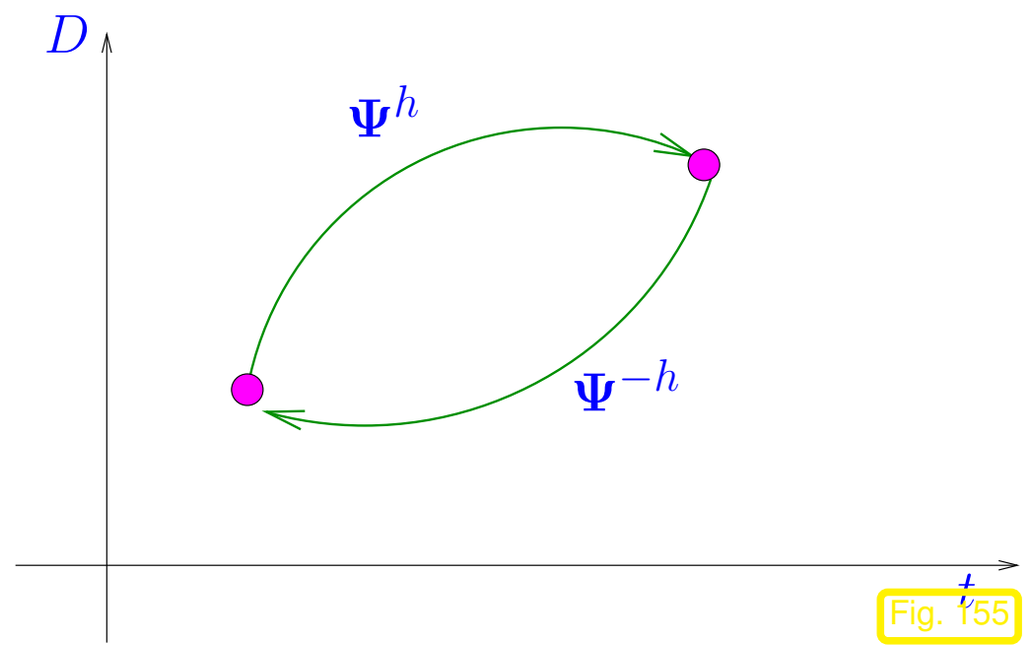


Fig. 155

Reminder of Thm. 2.1.29: reversible OSM have *even* order of consistence

We already know a reversible one-step method: implicit midpoint-rule (1.4.19) from Section 1.4.3, the simplest Gauss-collocation method, cf. (2.2.19).

Example 4.3.1 (Einfache symmetrische RK-ESV).

- Implicit midpoint-rule (1.4.19): $\mathbf{y}_1 = \mathbf{y}_0 + hf((\mathbf{y}_0 + \mathbf{y}_1)/2)$
- Implicit trapezoidal-rule: $\mathbf{y}_1 = \mathbf{y}_0 + \frac{h}{2}(f(\mathbf{y}_0) + f(\mathbf{y}_1))$

$\diamond$

Theorem 4.3.2 (Consistency order of symmetric OSM)**.**
The maximal consistency order ($\rightarrow$ Def. 2.1.13) of a symmetric one-step-method is even.

Theorem 4.3.3 (Reversible Runge-Kutta-one-step method)**.**
A s-level RK-OSM ($\rightarrow$ Def. 2.3.5) with Butcher-Tableau $\begin{array}{c|c} \mathbf{c} & \mathfrak{A} \\ \hline & \mathbf{b}^T \end{array}$*, see* (2.3.6)*, is reversible (symmetric, $\rightarrow$ Def. 2.1.27), if*

$$a_{s+1-i,s+1-j} + a_{ij} = b_j \quad \forall 1 \leq i,j \leq s \ .$$

Proof (see [21, Sect. V.2, Thm. 2.3])

to show: $\mathbf{y}_0 \xrightarrow{\mathbf{\Psi}^h} \mathbf{y}_1 \xrightarrow{\mathbf{\Psi}^{-h}} \mathbf{y}_0$.

Test invariance of the equation by permutation $\mathbf{y}_0 \leftrightarrow \mathbf{y}_1$, $h \leftrightarrow -h$ in the equation

$$\begin{cases} \mathbf{k}_i = \mathbf{f}(\mathbf{y}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j) \ , \\ \mathbf{y}_1 = \mathbf{y}_0 + h\sum_{i=1}^{s} b_i\mathbf{k}_i \ . \end{cases} \Rightarrow \begin{cases} \mathbf{k}_i = \mathbf{f}(\mathbf{y}_1 - h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j) \ , \\ \mathbf{y}_0 = \mathbf{y}_1 - h\sum_{i=1}^{s} b_i\mathbf{k}_i \ . \end{cases} \tag{4.3.4}$$

$$\Rightarrow \begin{cases} \mathbf{k}_i = \mathbf{f}(\mathbf{y}_0 + h\sum_{j=1}^{s} (b_j - a_{ij})\mathbf{k}_j) \ , \\ \mathbf{y}_1 = \mathbf{y}_0 + h\sum_{i=1}^{s} b_i\mathbf{k}_i \ . \end{cases} \tag{4.3.5}$$

$2a_{ij} = b_j$ $\Rightarrow$ equality after permutation $\mathbf{y}_0 \leftrightarrow \mathbf{y}_1$, $h \leftrightarrow -h$, unfortunately this does not yield useful RK-OS (Exception: implicit midpoint-rule (1.4.19) with $s = 1$, $a_{11} = \frac{1}{2}$, $b_1 = 1$)

! Note: $a_{s+1-i,s+1-j} + a_{ij} = b_j \quad \Rightarrow \quad b_{s+1-i} = b_i$

➢ Change of indices $i \leftarrow s+1-i$, $j \leftarrow s+1-j$ unter the assumption $b_{s+1-i} = b_i$

$$(4.3.5) \quad \Rightarrow \quad \begin{cases} \mathbf{k}_i = \mathbf{f}(\mathbf{y}_0 + h\sum_{j=1}^{s}(b_j - a_{s+1-i,s+1-j})\mathbf{k}_j) \ , \\ \\ \mathbf{y}_1 = \mathbf{y}_0 + h\sum_{i=1}^{s} b_i\mathbf{k}_i \ . \end{cases}$$

= Starting RK-OSM, if $b_j - a_{s+1-i,s+1-j} = a_{ij}$ □

Theorem 4.3.6 (Reversible Gauss-Collocation-OSM)**.**

Gauss-Collocation-OSM are reversible (→ Def. 2.1.27).

Proof. Since Gauss-Collocation-methods belong to the class of Runge-Kutta-one-step-method is is enough to verify the assumptions of Thm. 4.3.3. We use the explicit formula (2.2.3) for the Runge-

Kutta-coefficients a_{ij} and b_i, $1 \le i,j \le s$.

$$a_{ij} = \int_0^{c_i} L_j(\tau)\,\mathrm{d}\tau \quad , \quad b_i = \int_0^1 L_i(\tau)\,\mathrm{d}\tau \ ,$$

where c_i are the Collocations-points (= Gausspoints) normalized on $[0,1]$, and L_j are the corresponding Lagrange-polynomials, see (2.2.2).

Position of the Gauss-points for the s-point Gauss-quadrature-formula on $[0,1]$ is symmetric around $\frac{1}{2}$, see Fig. **??**:

$$c_i = 1 - c_{s+1-s} \quad \Rightarrow \quad L_i(\tau) = L_{s+1-i}(1-\tau) \ , \quad 1 \le i \le s \ . \tag{4.3.7}$$

$$\begin{aligned} a_{s+1-i,s+1-j} + a_{ij} &= \int_0^{c_{s+1-i}} L_{s+1-j}(\tau)\,\mathrm{d}\tau + \int_0^{c_i} L_j(\tau)\,\mathrm{d}\tau \\ &= -\int_1^{1-c_{s+1-i}} L_{s+1-j}(1-\tau)\,\mathrm{d}\tau + \int_0^{c_i} L_j(\tau)\,\mathrm{d}\tau \\ &\overset{(4.3.7)}{=} \int_{c_i}^1 L_j(\tau)\,\mathrm{d}\tau + \int_0^{c_i} L_j(\tau)\,\mathrm{d}\tau = b_j \ . \end{aligned}$$

Thm. 4.3.3 $\Rightarrow$ finished. □

Supplement to Sect. 3.3:

Theorem 4.3.8 (stability domain and Reversibility)**.**
For reversible and A-stable (→ Def. 3.2.16) Runge-Kutta-one-step-method we have $\mathcal{S}_{\mathbf{\Psi}} = \mathbb{C}^-$.

New concept: R-Reversibility = "generalized time-reversal-symmetry"

Example 4.3.9 (Reversibility of mechanical systems)*.*

Hamiltonian differential equation (→ Def. 1.2.20) with Hamiltonian-function

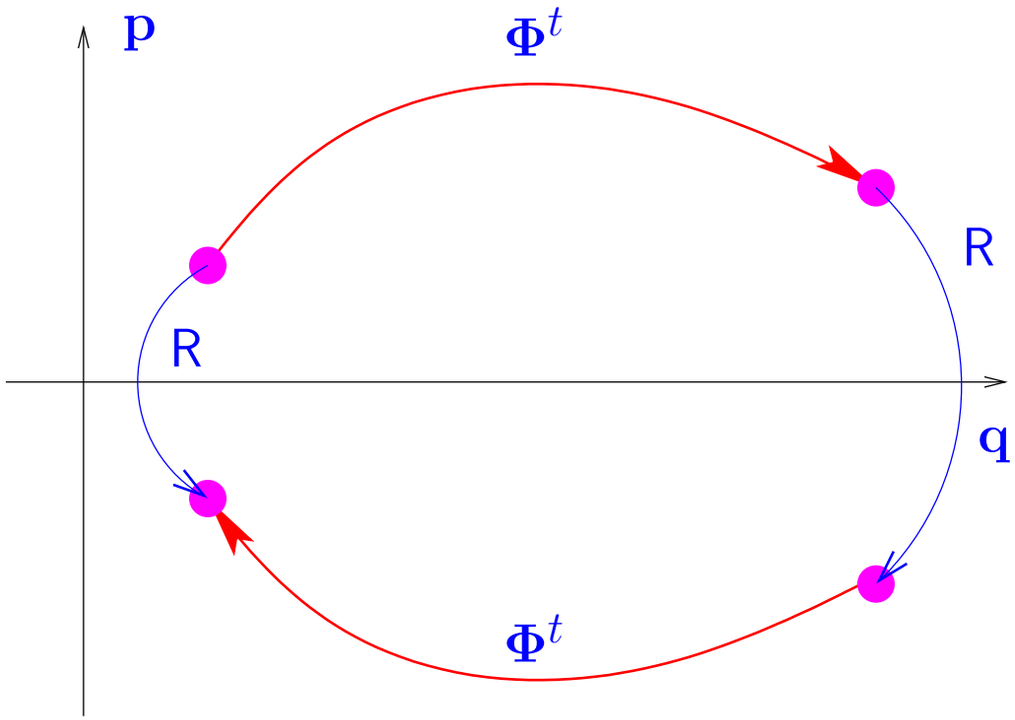
$$H : \begin{cases} \mathbb{R}^n \times \mathbb{R}^n & \mapsto \mathbb{R} \\ (\mathbf{p}, \mathbf{q}) & \mapsto \frac{1}{2}\mathbf{p}^T \mathbf{M}^{-1}\mathbf{p} + U(\mathbf{q}) \end{cases}, \quad \Rightarrow \quad H(\mathbf{p}, \mathbf{q}) = H(-\mathbf{p}, \mathbf{q}) \,,$$

with s.p.d. mass matrix $\mathbf{M} \in \mathbb{R}^{d,d}$.

$$\dot{\mathbf{p}}(t) = -\frac{\partial H}{\partial \mathbf{q}}(\mathbf{p}(t), \mathbf{q}(t)) = -\,\mathbf{grad}\, U(\mathbf{q}) \,, \dot{\mathbf{q}}(t) = \frac{\partial H}{\partial \mathbf{p}}(\mathbf{p}(t), \mathbf{q}(t)) = \mathbf{M}^{-1}\mathbf{p} \,. \tag{4.3.10}$$

Commutative diagram (Time-reversal-symmetry)

$$\begin{tikzcd}
(\mathbf{p}_0, \mathbf{q}_0) \arrow[r, "\text{evolution (4.3.10)}", "\text{from } 0 \text{ to } T"'] \arrow[d, "\text{Reversal of velocities: } \mathbf{p} \leftarrow -\mathbf{p}"'] & (\mathbf{p}(T), \mathbf{q}(T)) \arrow[d, "\text{Reversal of velocities: } \mathbf{p} \leftarrow -\mathbf{p}"] \\
(-\mathbf{p}_0, \mathbf{q}_0) & (-\mathbf{p}(T), \mathbf{q}(T)) \arrow[l, "\text{Evolution (4.3.10)}"', "\text{von } T \text{ bis } 0"]
\end{tikzcd}$$



$\Leftrightarrow$ evolution $\mathbf{\Phi}^t$ of (1.2.21) satisfies

$$\mathsf{R} \circ \mathbf{\Phi}^t = \mathbf{\Phi}^{-t} \circ \mathsf{R} \tag{4.3.11}$$

with map

$$\mathsf{R} \begin{pmatrix} \mathbf{p} \\ \mathbf{q} \end{pmatrix} = \begin{pmatrix} -\mathbf{p} \\ \mathbf{q} \end{pmatrix} . \tag{4.3.12}$$

(4.3.11) $\hat{=}$ "backward evolution" after reversal of velocities

$\diamond$

Abstraction:

Assume autonomous IVPs $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}), \quad \mathbf{y}(0) = \mathbf{y}_0,$

$\mathbf{f} : D \mapsto \mathbb{R}^d$ local Lipschitz-continuous ($\rightarrow$ Def. 1.3.2) **Assumption**: For all $\mathbf{y}_0 \in D$ exists the solution for all times, cf. Def. 1.3.1

Definition 4.3.13 (R-reversible map)**.**
Let $\mathsf{R} : D \mapsto D \subset \mathbb{R}^d$ *be a bijective* linear *map.*
Another bijective map $\Phi : D \mapsto D$ *is called* R*-reversible, if*

$$\mathsf{R} \circ \Phi = \Phi^{-1} \circ \mathsf{R} \,.$$

Lemma 4.3.14 (R-reversible evolutions)**.**
The Evolution $\mathbf{\Phi}^t$ *of* $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ *is* R*-reversible for all* $t \in \mathbb{R}$*, if*

$$\mathbf{f} \circ \mathsf{R} = -\mathsf{R} \circ \mathbf{f} \quad \text{on } D \,. \tag{4.3.15}$$

Proof. (see [21, Sect. V.1]) To show, because of $\mathbf{\Phi}^t \circ \mathbf{\Phi}^{-t} = Id$ (group property (1.3.8))

$$\mathsf{R} \circ \mathbf{\Phi}^t = (\mathbf{\Phi}^t)^{-1} \circ \mathsf{R} = \mathbf{\Phi}^{-t} \circ \mathsf{R} \,. \tag{4.3.16}$$

Idea: both sides of (4.3.16) are solutions of the same Initial Value Problem $\mathbf{y} \in D$

$$\frac{d}{dt}((\mathsf{R} \circ \mathbf{\Phi}^t)(\mathbf{y})) = \mathsf{R}\mathbf{f}(\mathbf{\Phi}^t(\mathbf{y})) = -\mathbf{f}((\mathsf{R} \circ \mathbf{\Phi}^t)(\mathbf{y})) \; , \tag{4.3.17}$$

$$\frac{d}{dt}((\mathbf{\Phi}^{-t} \circ \mathsf{R})(\mathbf{y})) = -\mathbf{f}((\mathbf{\Phi}^{-t} \circ \mathsf{R})(\mathbf{y})) \; . \tag{4.3.18}$$

$t \mapsto (\mathsf{R} \circ \mathbf{\Phi}^t)(\mathbf{y})$ and $t \mapsto (\mathbf{\Phi}^{-t} \circ \mathsf{R})(\mathbf{y})$ are both solutions of the initial value problem

$$\dot{\mathbf{z}} = -\mathbf{f}(\mathbf{z}) \quad , \quad \mathbf{z}(0) = \mathsf{R}\mathbf{y} \; .$$

Hence (4.3.16) follows from the uniqueness Theorem . 1.3.4. □

P. Grohs

rev 63606, February 17, 2014

Example 4.3.19 (Continuation: Reversibility of mechanical systems). Ex. 4.3.9

For hamiltonian evolution (4.3.10) with $\mathbf{y} = (\mathbf{p}, \mathbf{q})^T$, $d = 2n$, R on (4.3.12)

$$(\mathbf{f} \circ \mathsf{R})(\mathbf{y}) = \begin{pmatrix} -\operatorname{\mathbf{grad}} U(\mathsf{R}_{\mathbf{q}}(\mathbf{y})) \\ \mathbf{M}^{-1}\mathsf{R}_{\mathbf{p}}(\mathbf{y}) \end{pmatrix} = \begin{pmatrix} -\operatorname{\mathbf{grad}} U(\mathbf{q}) \\ -\mathbf{M}^{-1}\mathbf{p} \end{pmatrix} = -\mathsf{R}\begin{pmatrix} -\operatorname{\mathbf{grad}} U(\mathbf{q}) \\ \mathbf{M}^{-1}\mathbf{p} \end{pmatrix} = -\mathsf{R}(\mathbf{f}(\mathbf{y})) \; .$$

$\hat{=}$ requirement on lemma 4.3.14.

Alternative perspective: hamiltonian Dgl. (1.2.24) $\dot{\mathbf{y}} = \mathbf{J}^{-1}\,\mathbf{grad}\,H(\mathbf{y})$, $\mathbf{J} = \begin{pmatrix} 0 & \mathbf{I} \\ -\mathbf{I} & 0 \end{pmatrix}$:

$$H(\mathsf{R}\mathbf{y}) = H(\mathbf{y}) \quad \Rightarrow \quad \mathsf{R}\,\mathbf{grad}\,H(\mathsf{R}\mathbf{y}) = \mathbf{grad}\,H(\mathbf{y})\,. \tag{4.3.20}$$

For R from (4.3.11): $\mathbf{J} \circ \mathsf{R} = -\mathsf{R} \circ \mathbf{J}, \qquad \mathsf{R}^2 = Id$

$$\overset{(4.3.20)}{\Rightarrow} \quad -\mathsf{R}(\mathbf{J}^{-1}\,\mathbf{grad}\,H(\mathbf{y})) = \mathbf{J}^{-1}\mathsf{R}(\mathbf{grad}\,H(\mathbf{y})) = \mathbf{J}^{-1}\mathsf{R}\mathsf{R}\,\mathbf{grad}\,H(\mathsf{R}\mathbf{y}) = \mathbf{J}^{-1}\,\mathbf{grad}\,H(\mathsf{R}\mathbf{y})$$

$\hat{=}$ (4.3.15) for $\mathbf{f}(\mathbf{y}) = \mathbf{J}^{-1}\,\mathbf{grad}\,H(\mathbf{y})$.

Theorem 4.3.21 (R-reversible Runge-Kutta-evolutions)**.**
The right hand side $\mathbf{f}$ *of the autonomous ODE* $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ *satisfies* (4.3.15)*.*
Then the discrete evolution produced by a Runge-Kutta-one-step method (→ Def. 2.3.5) is R*-reversible, if and only if the RK-OSM is reversible/symmetric (→ Def. 2.1.27).*

Proof. (siehe [21, Sect. V.1, Thm. 1.5])

① With Notations of Lemma 4.3.14 and $\mathbf{\Psi}^h$ discrete evolution of the RK-OSM corresponding to the

P. Grohs

rev 63606, February 17, 2014

ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ we show that (cf. Proof of Affine-Covariance of RK-OSM, Rem. 2.3.13)

$$\mathbf{f} \circ \mathsf{R} = -\mathsf{R} \circ \mathbf{f} \quad \Rightarrow \quad \mathsf{R} \circ \boldsymbol{\Psi}^h = \boldsymbol{\Psi}^{-h} \circ \mathsf{R} \,. \tag{4.3.22}$$

By Def. 2.3.5, *linearity* of R

$$\begin{cases} \mathbf{k}_i = \mathbf{f}(\mathbf{y} + h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j) \,, \\ \boldsymbol{\Psi}^h\mathbf{y} = \mathbf{y} + h\sum_{i=1}^{s} b_i\mathbf{k}_i \,, \end{cases} \overset{(4.3.15)}{\Longrightarrow} \begin{cases} \mathsf{R}\mathbf{k}_i = -\mathbf{f}(\mathsf{R}\mathbf{y} + h\sum_{j=1}^{s} a_{ij}\mathsf{R}\mathbf{k}_j) \,, \\ \mathsf{R}\boldsymbol{\Psi}^h\mathbf{y} = \mathsf{R}\mathbf{y} + h\sum_{i=1}^{s} b_i\mathsf{R}\mathbf{k}_i \,. \end{cases}$$

- transformed increment $\widetilde{\mathbf{k}}_i := -\mathsf{R}\mathbf{k}_i$ satisfies

$$\widetilde{\mathbf{k}}_i = \mathbf{f}(\mathsf{R}\mathbf{y} - h\sum_{j=1}^{s} a_{ij}\widetilde{\mathbf{k}}_j) \,, \quad i = 1,\dots,s \,.$$

- $\widetilde{\mathbf{k}}_i \mathrel{\hat{=}}$ increments of the RK-OSM with stepwidth $-h$, Initial Value $\mathsf{R}\mathbf{y} \leftrightarrow \boldsymbol{\Psi}^{-h}\mathsf{R}\mathbf{y}$

$$\mathsf{R}\boldsymbol{\Psi}^h\mathbf{y} = \mathsf{R}\mathbf{y} - h\sum_{i=1}^{s} b_i\widetilde{\mathbf{k}}_i = \boldsymbol{\Psi}^{-h}\mathsf{R}\mathbf{y} \quad \Rightarrow \quad (4.3.22) \,.$$

② direct verification of Def. 4.3.13

RK-ESV reversible/symmetric

$$\Updownarrow \qquad\qquad \overset{(4.3.22)}{\Rightarrow} \qquad \mathsf{R} \circ \boldsymbol{\Psi}^h = \boldsymbol{\Psi}^{-h} \circ \mathsf{R} = (\boldsymbol{\Psi}^h)^{-1} \circ \mathsf{R} \,. \qquad \square$$

$$\boldsymbol{\Psi}^{-h} = (\boldsymbol{\Psi}^h)^{-1}$$

4.4 Symplecticity

4.4.1 Symplectic evolutions of hamiltonian differential equations

Reminder (Sect. 1.2.4): Hamiltonian differential equation $\rightarrow$ Def. 1.2.20

$$\dot{\mathbf{p}}(t) = -\frac{\partial H}{\partial \mathbf{q}}(\mathbf{p}(t),\mathbf{q}(t)) \quad , \quad \dot{\mathbf{q}}(t) = \frac{\partial H}{\partial \mathbf{p}}(\mathbf{p}(t),\mathbf{q}(t)) \; , \tag{1.2.21}$$

with (smooth) Hamiltonian-function $H : \mathbb{R}^n \times M \mapsto \mathbb{R}$, Configuration space $M \subset \mathbb{R}^n$.

$$\mathbf{y} = \begin{pmatrix}\mathbf{p}\\ \mathbf{q}\end{pmatrix} \quad \Longrightarrow \quad (1.2.21) \quad \Leftrightarrow \quad \boxed{\dot{\mathbf{y}} = \mathbf{J}^{-1} \cdot \mathbf{grad}\, H(\mathbf{y})} \; , \quad \mathbf{J} = \begin{pmatrix} 0 & \mathbf{I}_n \\ -\mathbf{I}_n & 0 \end{pmatrix} \in \mathbb{R}^{2n,2n} \; . \tag{1.2.24}$$

Lemma 1.2.23 (Conservation of energy): H is Invariant of (1.2.21)

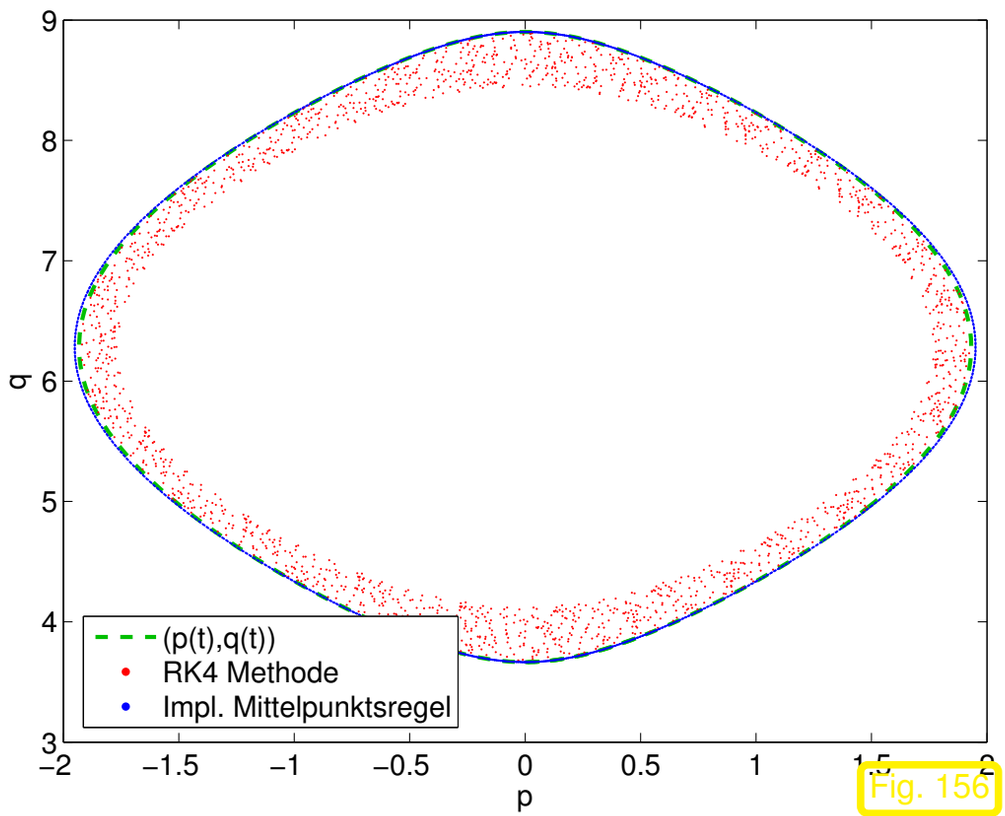
P. Grohs

rev 63606, February 17, 2014

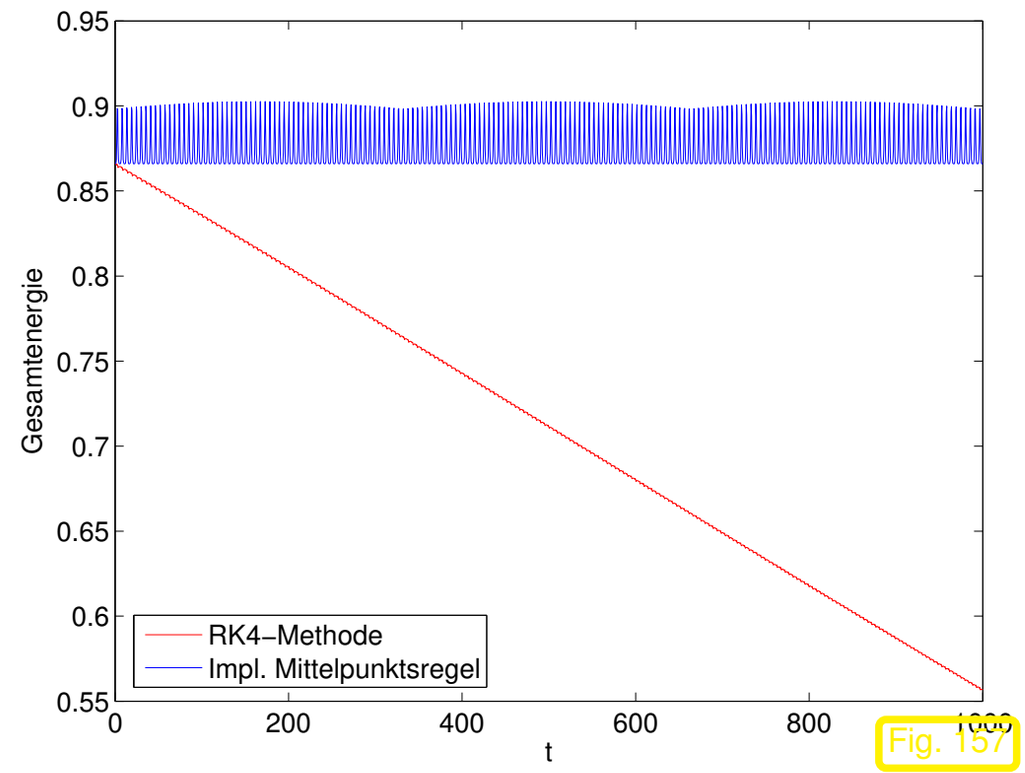
Example 4.4.1 (Conservation of energy at numerical integration). $\leftrightarrow$ Ex. 1.4.24

mathematical pendulum Ex. 1.2.17, IVP for (1.2.19) on $[0, 1000]$, $p(0) = 0$, $q(0) = 7\pi/6$.

Comparison to classical Runge-Kutta-method (2.3.11) (order 4) with 1-level Gauss-Collocations-OSM (implicit midpoint-rule 2.2.19), equidistant mesh, $h = \frac{1}{2}$:



Trajections of "exact" /discrete evolutions



Energy conservation of discrete evolutions

No Energy shift with implicit midpoint-rule

A mysterious observation:

Characteristic of some $(*)$ numerical integrators:

Approximative longtime-energyconservation (no Energy shift)

implicit midpoint-rule (1.4.19) $\rightarrow$ Ex. 4.4.1, 1.4.24,

Störmer-Verlet-method (2.5.13) $\rightarrow$ Ex. 1.4.32

Remark 4.4.2 (Conservation of volume of two-dimensional hamiltonian ODEs).

For Evolution $\mathbf{\Phi}^t : \mathbb{R}^n \times M \mapsto \mathbb{R}^n \times M$ of a hamiltonian differential equation we have:

$$n = 1 \quad \succ \quad \operatorname{div}_{\mathbf{y}} \underbrace{\mathbf{J}^{-1} \mathbf{grad}\, H(\mathbf{y})}_{\operatorname{rot} H(\mathbf{y})} = 0 \quad \overset{\text{Thm. 4.2.3}}{\succ} \quad \mathbf{\Phi}^t \text{ conservation of volume (area).}$$

△

Example 4.4.3 (Conservation of area of Evolution for pendulum equation). $\rightarrow$ Ex. 1.2.17

$p \leftrightarrow$ angular velocity, $q \leftrightarrow$ angular variable α

$$\begin{aligned} \dot{p} &= -\sin q \;, \\ \dot{q} &= p \end{aligned}$$

$\blacktriangleright$ hamiltonian-function $H(p,q) = \frac{1}{2}p^2 - \cos q$ (total energy) (4.4.4)

Conservation of volume in state space:

Evolution of a quadratic volume $\triangleright$

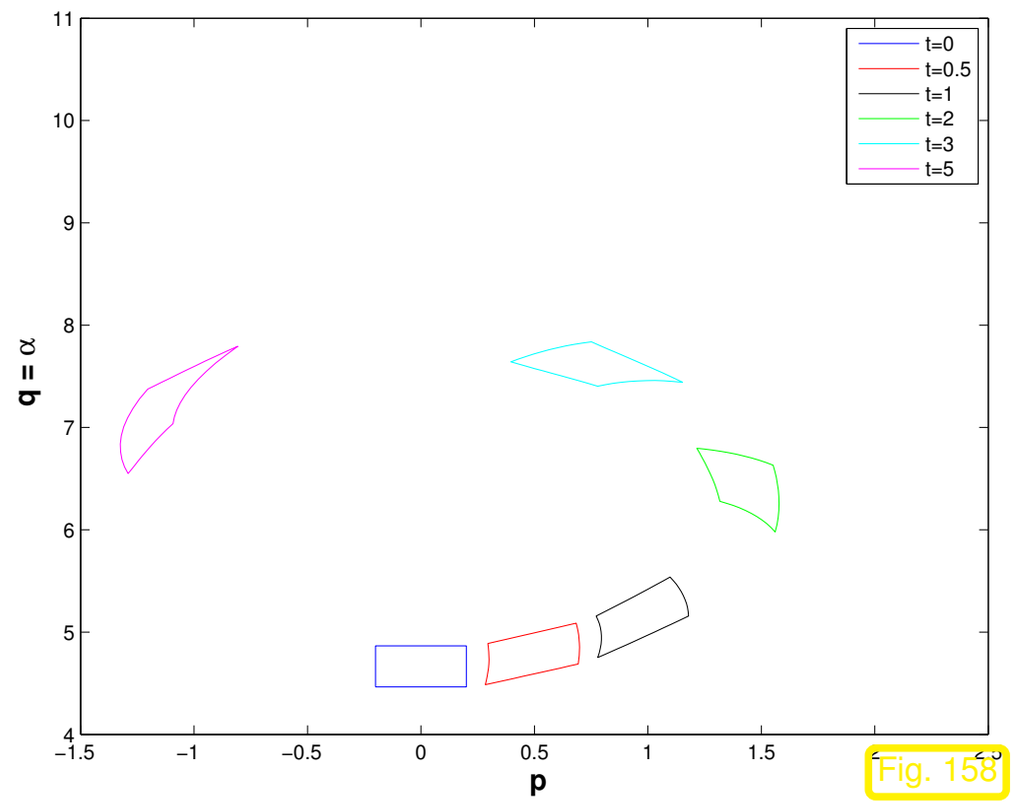
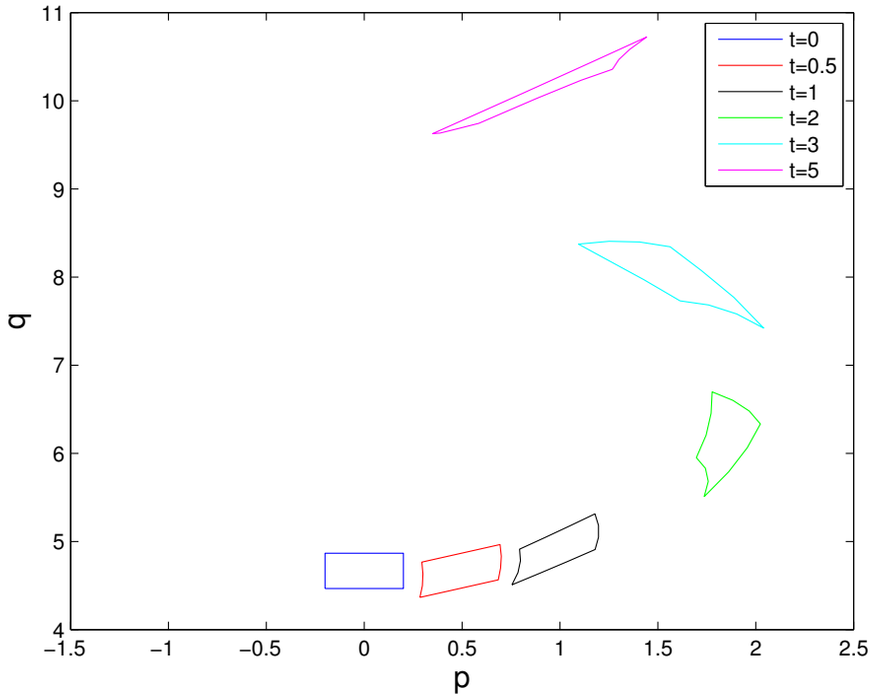


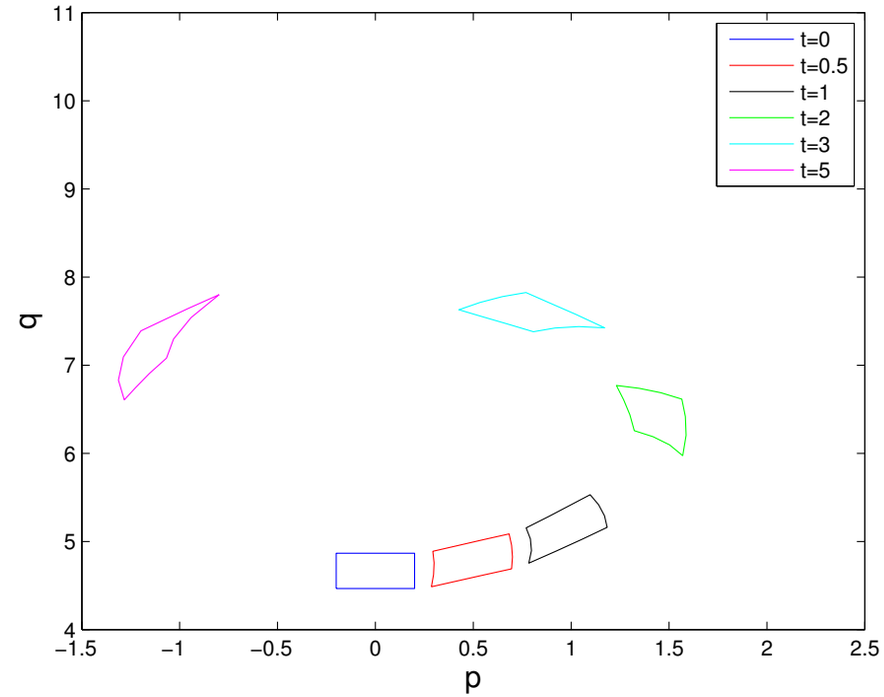
Fig. 158

P. Grohs

rev 63606, February 17, 2014



Evolution of a quadratic volume (explicit eulermethod)



evolution of a quadratic volume (implicit midpoint-rule)

P. Grohs

rev 63606, February 17, 2014

Rem. 4.2.9: For $d = 2$ the implicit midpoint-rule conservation of volume (like all gauss-collocation-methods by Lemma 4.1.6)

◇

Generalization of higher dimensions: Symplectic evolutions ($\neq$ conservation of volume)

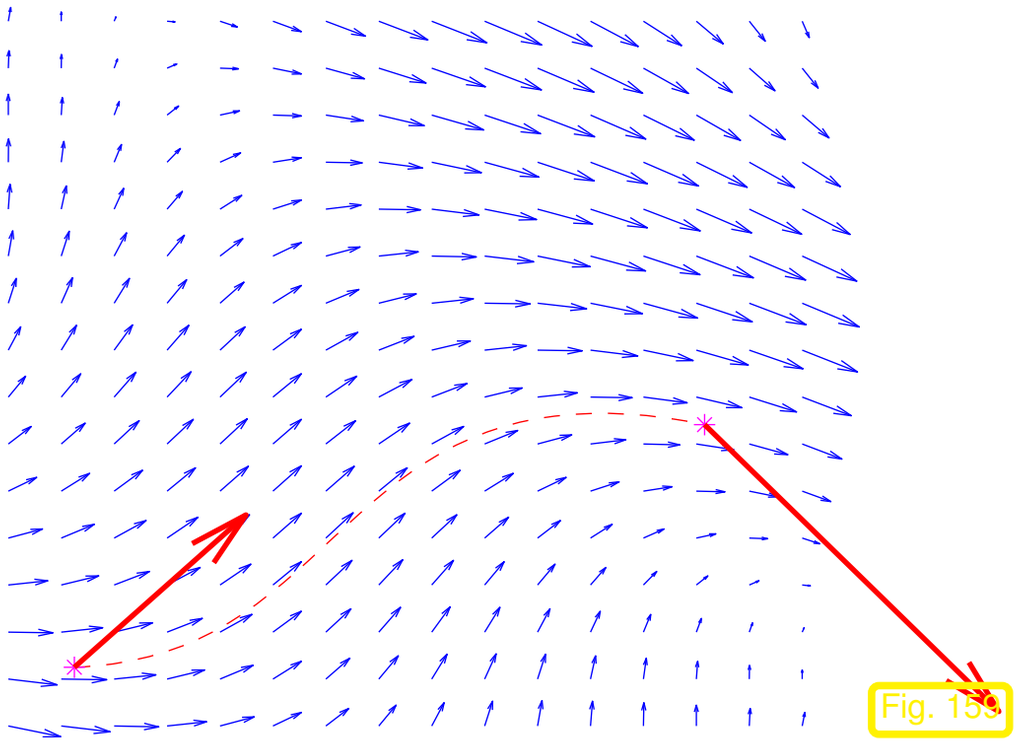


Fig. 159

Push-Forward: Effect of a (smooth) map on infinitesimal length = Vector for C^1-map $\mathbf{\Phi} : D \subset \mathbb{R}^d \mapsto \mathbb{R}^d$:

$$(\mathbf{\Phi}_* \mathbf{v})(\mathbf{y}) = D\mathbf{\Phi}(\mathbf{y})\mathbf{v} \quad \mathbf{y} \in D, \mathbf{v} \in \mathbb{R}^d \,.$$

Transport of a vector in "flow field" $t \mapsto \mathbf{\Phi}^t$ to $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$

Definition 4.4.5 (symplectic product)**.**

$$\omega(\mathbf{v}, \mathbf{w}) := \mathbf{v}^T \mathbf{J} \mathbf{w} \,, \quad \mathbf{v}, \mathbf{w} \in \mathbb{R}^{2n} \quad \textit{mit} \quad \mathbf{J} = \begin{pmatrix} 0 & \mathbf{I}_n \\ -\mathbf{I}_n & 0 \end{pmatrix} \,.$$

Remark 4.4.6 (constant 2-forms).

symplectic product $\hat{=}$ prototype of a nondegenerated, alternating bilinearform:

Definition 4.4.7 (alternating, nondegenerated bilinearform).
A bilinearform $\beta : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ *is called*

- *alternating* $:\Leftrightarrow \quad \beta(\mathbf{x},\mathbf{y}) = -\beta(\mathbf{y},\mathbf{x}) \quad \forall \mathbf{x},\mathbf{y} \in \mathbb{R}^d$,
- *nondegenerated* $:\Leftrightarrow \quad \beta(\mathbf{x},\mathbf{y}) = 0 \quad \forall \mathbf{y} \in \mathbb{R}^d \quad \Rightarrow \quad \mathbf{x} = 0$

$\beta : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ alternating bilinearform $\quad \Rightarrow \quad \exists \mathbf{L} \in \mathbb{R}^{d,d}$: $\quad \mathbf{L}^T = -\mathbf{L}$

$$\beta(\mathbf{x},\mathbf{y}) = \mathbf{x}^T\mathbf{L}\mathbf{y} \quad \forall \mathbf{x},\mathbf{y} \in \mathbb{R}^d$$

Lemma 4.4.8 (normalform of skew-symmetric matrices).
For each regular $\mathbf{L} \in \mathbb{R}^{2n,2n}$ *with* $\mathbf{L}^T = -\mathbf{L}$ *there exists a regular* $\mathbf{U} \in \mathbb{R}^{2n,2n}$, *such that*

$$\mathbf{U}^T\mathbf{L}\mathbf{U} = \mathbf{J} = \begin{pmatrix} 0 & \mathbf{I}_n \\ -\mathbf{I}_n & 0 \end{pmatrix} \qquad \textit{(Congruencetransformation).}$$

P. Grohs

rev 63606, February 17, 2014

Proof. $\mathbf{L} = -\mathbf{L}^T \quad \Rightarrow$ unitary diagonalized (normal matrix !), proper imaginary eigenvalues, which occur in complex conjugated pairs with complex conjugated eigenvectors:

$$\exists \mathbf{Q} \in \mathbb{C}^{2n}: \quad \mathbf{Q}^{-1} = \mathbf{Q}^H \quad \text{and} \quad \mathbf{Q}^H \mathbf{L} \mathbf{Q} = i \begin{pmatrix} \mathbf{D} & 0 \\ 0 & -\mathbf{D} \end{pmatrix} ,$$

with $\mathbf{D} = \operatorname{diag}(\mu_1, \dots, \mu_n) \in \mathbb{R}^n$, $\mu_i > 0$. Then put

$$\mathbf{U} = \frac{1}{\sqrt{2}} \mathbf{Q} \begin{pmatrix} \mathbf{D}^{-1/2} & \mathbf{D}^{-1/2} \\ -i\mathbf{D}^{-1/2} & i\mathbf{D}^{-1/2} \end{pmatrix} . \qquad \square$$

Note: The matrix $\mathbf{U}$ is real !

There is a real coordinate transformation from β to ω ($\rightarrow$ Def. 4.4.5). $\triangle$

P. Grohs

rev 63606, February 17, 2014

Remark 4.4.9 (Symplectic flowintegral).

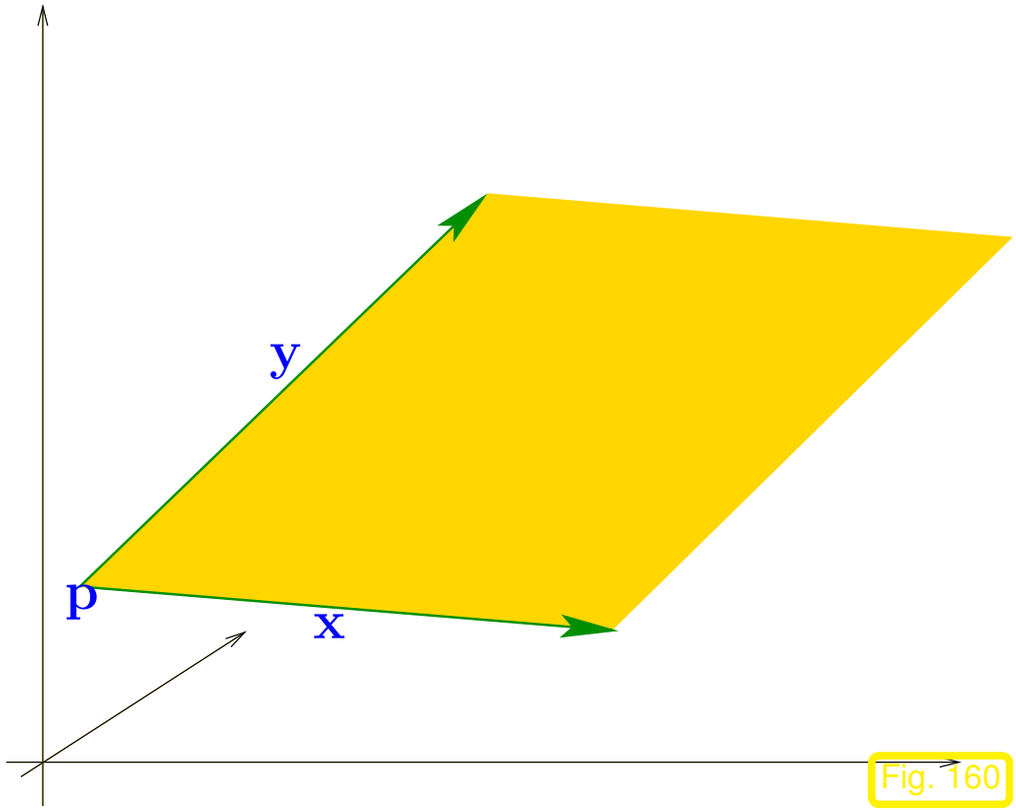


Fig. 160

$\omega(\mathbf{x},\mathbf{y}) \mathrel{\hat{=}}$ Flow "through" oriented parallelogram, spanned by $\{\mathbf{p}, \mathbf{p}+\mathbf{x}, \mathbf{p}+\mathbf{y}, \mathbf{p}+\mathbf{x}+\mathbf{y}\}$ (weighted area)

"Riemann-summation" ▷

Flow through bounded oriented differentiated area (= manifold of dimension 2)

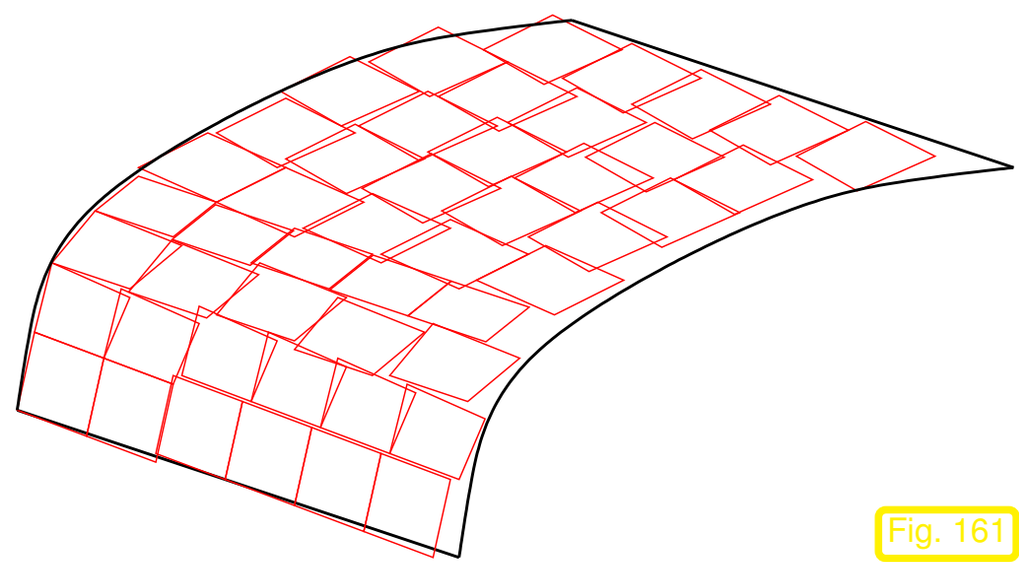


Fig. 161

P. Grohs

rev 63606, February 17, 2014

If $\psi : U \mapsto \mathbb{R}^d$ is a parametrization (map) of a 2-manifold Σ, we have, cf. Push-Forward,

$$\text{Flow} = \int_\Sigma \omega = \int_U \left(\frac{d\psi}{du_1}\right)^T \mathbf{J} \left(\frac{d\psi}{du_2}\right) \, \mathrm{d}\mathbf{u} \; . \tag{4.4.10}$$

△

Theorem 4.4.11 (Symplectic Flow of hamiltonian systems)**.**
Let $\mathbf{\Phi}^t$ be the Evolution of a hamiltonian differential equation (**??**) *with C^2-hamiltonian-function $H : \mathbb{R}^n \times M \mapsto \mathbb{R}$. Then we have*

$$\forall \mathbf{y} \in D\colon \quad \exists \delta > 0\colon \quad \omega\big((\mathbf{\Phi}^t_* \mathbf{v})(\mathbf{y}), (\mathbf{\Phi}^t_* \mathbf{w})(\mathbf{y})\big) = \omega(\mathbf{v}, \mathbf{w}) \quad \forall \mathbf{v}, \mathbf{w} \in \mathbb{R}^{2n}, \; 0 \le t < \delta \; .$$

P. Grohs

rev 63606, February 17, 2014

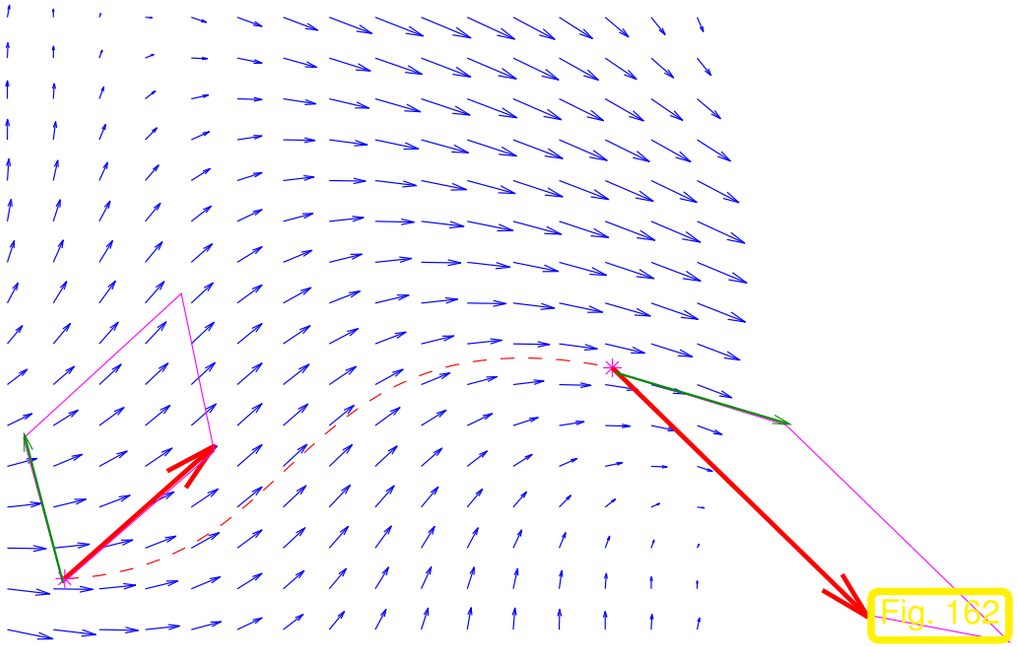


Illustration Push-Forward of two vectors and alternating (area) bilinearform.

Proof. ($\rightarrow$ Proof of [21, Thm. 2.4, Ch. VI])

$\mathbf{\Phi}^t \hat{=}$ evolution operator of the hamiltonian ODE $\dot{\mathbf{y}} = \mathbf{J}^{-1}\,\mathbf{grad}\,H(\mathbf{y})$

$$\text{Claim} \iff (\mathbf{\Phi}^t_*(\mathbf{v})\mathbf{y})^T\mathbf{J}(\mathbf{\Phi}^t_*(\mathbf{w})\mathbf{y}) = \mathbf{v}^T\mathbf{J}\mathbf{w} \quad \forall \mathbf{v},\mathbf{w} \in \mathbb{R}^d,\ \forall \mathbf{y} \in D\ ,$$

$$\iff \left(\frac{d}{d\mathbf{y}}\mathbf{\Phi}^t(\mathbf{y})\right)^T \mathbf{J}\left(\frac{d}{d\mathbf{y}}\mathbf{\Phi}^t(\mathbf{y})\right) = \mathbf{J} \quad \forall \mathbf{y} \in D\ .$$

propagationmatrix $\mathbf{W}(t;\mathbf{y}) := \frac{d}{d\mathbf{y}}\mathbf{\Phi}^t\mathbf{y}$ does solve variational equation (1.3.34)

$$\dot{\mathbf{W}}(t;\mathbf{y}) = D(\mathbf{J}^{-1}\,\mathbf{grad}\,H(\mathbf{y}))\mathbf{W}(t;\mathbf{y}) = \mathbf{J}^{-1}\mathbf{\nabla}^2 H(\mathbf{y})\mathbf{W}(t;\mathbf{y})\ , \quad \mathbf{y} \in D\ .$$

P. Grohs

rev 63606, February 17, 2014

✎ notation: $\nabla^2 H \hat{=}$ (symmetric) Hesse-Matrix of the Hamiltonian-function H.

With product rule, since $\mathbf{J}^T = -\mathbf{J}$, $\mathbf{J}^{-T} = -\mathbf{J}^{-1} = \mathbf{J}$:

$$\frac{d}{dt}\Big(\mathbf{W}(t;\mathbf{y})^T\mathbf{J}\mathbf{W}(t;\mathbf{y})\Big) = \dot{\mathbf{W}}(t;\mathbf{y})^T\mathbf{J}\mathbf{W}(t;\mathbf{y}) + \mathbf{W}(t;\mathbf{y})^T\mathbf{J}\dot{\mathbf{W}}(t;\mathbf{y})$$
$$= \mathbf{W}(t;\mathbf{y})^T\nabla^2 H(\mathbf{y})\underbrace{\mathbf{J}^{-T}\mathbf{J}}_{=-\mathbf{I}}\mathbf{W}(t;\mathbf{y}) + \mathbf{W}(t;\mathbf{y})^T\underbrace{\mathbf{J}\mathbf{J}^{-1}}_{=\mathbf{I}}\nabla^2 H(\mathbf{y})\mathbf{W}(t;\mathbf{y}) = 0\,.$$

As $\mathbf{W}(0;\mathbf{y}) = \mathbf{I} \quad\Rightarrow\quad \mathbf{W}(t;\mathbf{y})^T\mathbf{J}\mathbf{W}(t;\mathbf{y}) = \mathbf{J} \quad \forall t$ □

Definition 4.4.12 (symplectic map)**.**
A C^1-map $\mathbf{\Phi} : D \subset \mathbb{R}^{2n} \mapsto \mathbb{R}^{2n}$ is called symplectic, if

$$D\mathbf{\Phi}(\mathbf{y})^T\mathbf{J}D\mathbf{\Phi}(\mathbf{y}) = \mathbf{J} \quad\Leftrightarrow\quad \omega\big(\underbrace{D\mathbf{\Phi}(\mathbf{y})\mathbf{v}}_{(\mathbf{\Phi}_*\mathbf{v})(\mathbf{y})}, \underbrace{D\mathbf{\Phi}(\mathbf{y})\mathbf{w}}_{(\mathbf{\Phi}_*\mathbf{w})(\mathbf{y})}\big) = \omega(\mathbf{v},\mathbf{w}) \quad \forall \mathbf{v},\mathbf{w} \in \mathbb{R}^{2n}, \forall \mathbf{y} \in D\,.$$

Thm. 4.4.11: The Evolution to a hamiltonian differential equation is symplectic at every point in time.

P. Grohs

rev 63606, February 17, 2014

The Concept of Symplecticity is strongly connected with the differentialgeometrical view of hamiltonian evolution, see [3, Part III].

Corollary 4.4.13 (Composition of symplectical maps)**.**

The Composition of symplectical maps is symplectic.

Remark 4.4.14 (Vectorspaces of vektorfields and properties of evolutions)*.*

Def. 1.3.7: Vectorfield $\mathbf{f} : D \subset \mathbb{R}^d \mapsto \mathbb{R}^d$ ➢ Evolution $\mathbf{\Phi}^t$ to the ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$

(1.2.8):	$\mathbf{grad}\, I \cdot \mathbf{f} = 0$	$\Leftrightarrow$	$\mathbf{\Phi}^t$ "I-isoareaconserving" for all t
Thm. 4.2.3:	$\operatorname{div} \mathbf{f} = 0$	$\Leftrightarrow$	$\mathbf{\Phi}^t$ volume conserving ($\rightarrow$ Def. 4.2.1) $\forall t$
Lemma 4.3.14:	$\mathbf{f} \circ \mathsf{R} = -\mathsf{R} \circ \mathbf{f}$	$\Leftrightarrow$	$\mathbf{\Phi}^t$ R-reversible ($\rightarrow$ Def. 4.3.13) $\forall t$
Thm. 4.4.11:	$\mathbf{f} = \mathbf{J}^{-1}\mathbf{grad}\, H$	$\Rightarrow$	$\mathbf{\Phi}^t$ symplectic ($\rightarrow$ Def. 4.4.12) $\forall t$

vectorspace V of vectorfield $D \mapsto \mathbb{R}^d$ ▶ *group* $\mathfrak{G}$ of diffeomorphism

One-step-method for $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ structurpreserving $:\Leftrightarrow \mathbf{f} \in V \quad \Rightarrow \quad \mathbf{\Psi}^h \in \mathfrak{G}$
(with discrete Evolution $\mathbf{\Psi}^h$)

△

Theorem 4.4.15 (Symplectic Evolution and hamiltonian differential equation)**.**
Let $\mathbf{\Phi}^t$ be the evolutionoperator to a autonomous ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\mathbf{f} : D \subset \mathbb{R}^{2n} \mapsto \mathbb{R}^{2n}$ continuously differentiable, State space D star-shaped*. Then we have*

$$\mathbf{\Phi}^t \quad \textit{symplectic } (\rightarrow \textit{Def. 4.4.12}) \; \forall t \qquad \Leftrightarrow \quad \exists H : D \mapsto \mathbb{R}: \quad \mathbf{f}(\mathbf{y}) = \mathbf{J}^{-1} \mathbf{grad}\, H(\mathbf{y}) \; .$$

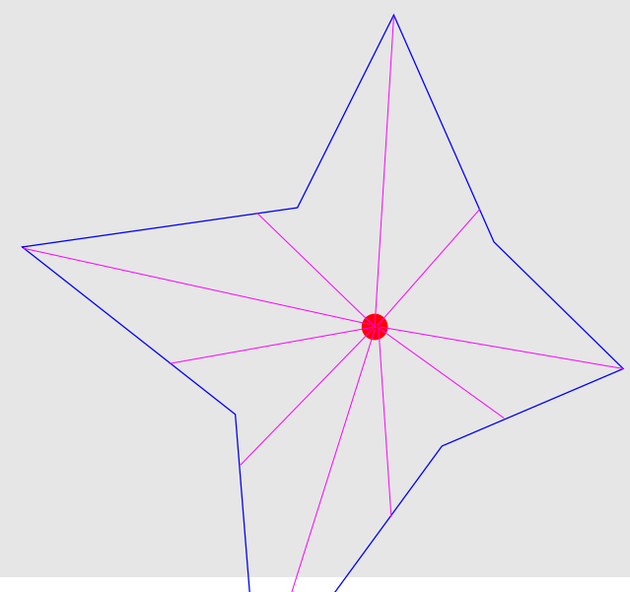
P. Grohs

rev 63606, February 17, 2014

Definition 4.4.16 (star-shaped domain)**.**
$D \subset \mathbb{R}^d$ *is called* star-shaped*, if there exists $\mathbf{z} \in D$, such that*

$$\{t\mathbf{z} + (1-t)\mathbf{x},\; 0 \le t \le 1\} \subset D$$

for all points $\mathbf{x} \in D$.



Utilities of the proof:

Lemma 4.4.17 (lemma of integrability)**.**
Let $D \subset \mathbb{R}^d$ be star-shaped and $\mathbf{f} : D \mapsto \mathbb{R}^d$ continuously differentiable. Then we have

$$D\mathbf{f} = D\mathbf{f}^T \quad \Leftrightarrow \quad \exists F : D \mapsto \mathbb{R}: \quad \mathbf{f}(\mathbf{y}) = \mathbf{grad}\, F(\mathbf{y}) \quad \forall \mathbf{y} \in D \; .$$

Proof. WLOG: D star-shaped wrt. 0 $\Rightarrow$ Well-defined is the function

$$F(\mathbf{y}) := \int_0^1 \mathbf{f}(\tau\mathbf{y}) \cdot \mathbf{y} \, \mathrm{d}\tau \quad \mathbf{y} \in D \; .$$

$$\Rightarrow \quad \mathbf{grad}\, F(\mathbf{y}) = DF(\mathbf{y})^T = \int_0^1 \tau D\mathbf{f}(\tau\mathbf{y})^T \cdot \mathbf{y} + \mathbf{f}(\tau\mathbf{y}) \, \mathrm{d}\tau = \int_0^1 \frac{d}{d\tau} \left(\mathbf{f}(\tau\mathbf{y})\tau\right)(\tau) \, \mathrm{d}\tau = \mathbf{f}(\mathbf{y}) \; .$$

Attention: strict differentiation of row- and columnvectors is important; gradient is a columnvector! □

Proof (of Thm. 4.4.15)

"$\Leftarrow$": See Thm. 4.4.11

"$\Rightarrow$": propagationmatrix $\mathbf{W}(t;\mathbf{y}) := (\frac{d}{d\mathbf{y}}\mathbf{\Phi}^t)(\mathbf{y})$ solves variational equation (1.3.34)

$$\dot{\mathbf{W}}(t;\mathbf{y}) = D\mathbf{f}(\mathbf{\Phi}^t\mathbf{y})\mathbf{W}(t;\mathbf{y}) \quad , \quad \mathbf{W}(0;\mathbf{y}) = \mathbf{I} \, , \quad \mathbf{y} \in D \, , \quad t \in J(\mathbf{y}) \, .$$

t fixes, sufficiently small: $\mathbf{y} \mapsto \mathbf{\Phi}^t\mathbf{y}$ is symplectic map ($\rightarrow$ Def. 4.4.12)

$$\blacktriangleright \qquad \mathbf{W}(t;\mathbf{y})^T\mathbf{J}\mathbf{W}(t;\mathbf{y}) = \mathbf{J} \quad \overset{t \text{ frei}}{\Longrightarrow} \quad \frac{d}{dt}\Big(\mathbf{W}(t;\mathbf{y})^T\mathbf{J}\mathbf{W}(t;\mathbf{y})\Big) = 0 \, .$$

With product rule:

$$\begin{aligned} 0 = \frac{d}{dt}\Big(\mathbf{W}(t;\mathbf{y})^T\mathbf{J}\mathbf{W}(t;\mathbf{y})\Big) &= \dot{\mathbf{W}}(t;\mathbf{y})^T\mathbf{J}\mathbf{W}(t;\mathbf{y}) + \mathbf{W}(t;\mathbf{y})^T\mathbf{J}\dot{\mathbf{W}}(t;\mathbf{y}) \\ &= (D\mathbf{f}(\mathbf{\Phi}^t\mathbf{y})\mathbf{W}(t;\mathbf{y}))^T\mathbf{J}\mathbf{W}(t;\mathbf{y}) + \mathbf{W}(t;\mathbf{y})^T\mathbf{J}(D\mathbf{f}(\mathbf{\Phi}^t\mathbf{y})\mathbf{W}(t;\mathbf{y})) \quad \forall \mathbf{y} \in D, \, |t| \text{ small.} \end{aligned}$$

Set $t = 0$, use $\mathbf{J}^{-T} = -\mathbf{J}^{-1} = \mathbf{J}$ $\qquad \Rightarrow \qquad \boxed{\mathbf{J}D\mathbf{f}(\mathbf{y}) = (\mathbf{J}D\mathbf{f}(\mathbf{y}))^T \quad \forall \mathbf{y} \in D}$

Because of $\mathbf{J}D\mathbf{f}(\mathbf{y}) = D(\mathbf{J}\mathbf{f})(\mathbf{y})$ Application of the integrabilitylemma 4.4.17. $\square$

4.4.2 Symplectic Integrators

Why is the numerical mathematician interested in this "exotic" property called "symplecticity"**?**

Thm. 4.4.15: $\mathbf{f} = \mathbf{J}^{-1}\,\mathbf{grad}\,H$ ("Equation of motion") $\Leftrightarrow$ $\mathbf{\Phi}^t$ symplectic ($\rightarrow$ Def. 4.4.12) $\forall t$

Intuition: discrete evolution $\mathbf{\Psi}^h$ symplectic $\leftrightarrow$ "discrete equation of motion

Symplecticity can be inherited by discrete evolutions!

P. Grohs

rev 63606, February 17, 2014

Definition 4.4.18 (Symplectic one-step-method)**.**
A one-step-method ($\rightarrow$ Def. 2.1.2) is called symplectic, if , applied to a hamiltonian differential equation ($\rightarrow$ Def. 1.2.20) $\dot{\mathbf{y}} = \mathbf{J}^{-1}\,\mathbf{grad}\,H(\mathbf{y})$ produced a consistent discrete Evolution $\mathbf{\Psi}^h$, such that $\mathbf{\Psi}^h : K \subset D \mapsto \mathbb{R}^d$ for each compact $K \subset D$ and constant sufficiently small $h > 0$ a symplectic map ($\rightarrow$ Def. 4.4.12).

Remark 4.4.19 (Simple symplectic integrators).

The discrete evolutions $\mathbf{\Psi}^h : D \subset \mathbb{R}^{2n} \mapsto \mathbb{R}^{2n}$ of the hamiltonian ODE (1.2.21) ($\dot{\mathbf{y}} = \mathbf{J}^{-1}\,\mathbf{grad}\,H(\mathbf{y})$, $H : D \subset \mathbb{R}^d \mapsto \mathbb{R}$) generated by

implicit midpoint-rule (1.4.19) symplectic eulermethod (2.5.11) Rem. 1.4.33, Ex. 2.5.10) for separated hamiltonian-function of the form $H(\mathbf{y}) = T(\mathbf{p}) + U(\mathbf{q})$, $\mathbf{y} = \binom{\mathbf{p}}{\mathbf{q}}$, are symplectic (for sufficiently small stepwidth $h \in \mathbb{R}$).

Verification of symplecticity:

- for implicit midpoint-rule (1.4.19):

$$\mathbf{\Psi}^h\mathbf{y}_0 := \mathbf{y}_1 = \mathbf{y}_0 + h\mathbf{J}^{-1}\,\mathbf{grad}\,H(\tfrac{1}{2}(\mathbf{y}_0 + \mathbf{y}_1))\ . \tag{4.4.20}$$

implicit differentiation (Assumption: H "sufficiently smooth"):

$$D\mathbf{\Psi}^h(\mathbf{y}_0) = \mathbf{I} + h\mathbf{J}^{-1}\mathbf{\nabla}^2 H(\tfrac{1}{2}(\mathbf{y}_0 + \mathbf{y}_1))\tfrac{1}{2}(\mathbf{I} + D\mathbf{\Phi}^h(\mathbf{y}_0))\ ,$$

$$\Rightarrow \quad D\mathbf{\Psi}^h(\mathbf{y}_0) = \left(\mathbf{I} - \tfrac{1}{2}h\mathbf{J}^{-1}\mathbf{\nabla}^2 H(\ldots)\right)^{-1}\left(\mathbf{I} + \tfrac{1}{2}h\mathbf{J}^{-1}\mathbf{\nabla}^2 H(\ldots)\right)\ .$$

Use ($\mathbf{M}$ sufficiently small)

$$\mathbf{M} = \mathbf{M}^T \quad \Rightarrow \quad (\mathbf{I} - \mathbf{JM})^T(\mathbf{I} + \mathbf{JM})^{-T}\mathbf{J}(\mathbf{I} + \mathbf{JM})^{-1}(\mathbf{I} - \mathbf{JM}) = \mathbf{J}\ . \tag{4.4.21}$$

- Störmer-Verlet-method (2.5.13) for $H(\mathbf{p},\mathbf{q}) = T(\mathbf{p}) + U(\mathbf{q})$:

$$\begin{cases} \mathbf{p}_{1/2} &= \mathbf{p}_0 - \frac{1}{2}h\,\mathbf{grad}\,U(\mathbf{q}_0)\;, \\ \mathbf{q}_1 &= \mathbf{q}_0 + h\,\mathbf{grad}\,T(\mathbf{p}_{1/2})\;, \\ \mathbf{p}_1 &= \mathbf{p}_{1/2} - \frac{1}{2}h\,\mathbf{grad}\,U(\mathbf{q}_1)\;. \end{cases} \tag{4.4.22}$$

Strang-Splittingmethod (Rem. 1.4.33): discrete evolution $\mathbf{\Psi}^h$ to (4.4.22) satisfies

$$\mathbf{\Psi}^h = \mathbf{\Phi}_U^{h/2} \circ \mathbf{\Phi}_T^h \circ \mathbf{\Phi}_U^{h/2}\;,$$

where $\mathbf{\Phi}_T^t$, $\mathbf{\Phi}_U^t$ *exact* evolutionoperator of hamiltonian ODE

$$\mathbf{\Phi}_T^t \quad \leftrightarrow \quad \begin{cases} \dot{\mathbf{p}} &= 0\;, \\ \dot{\mathbf{q}} &= \mathbf{grad}\,T(\mathbf{p}) \end{cases} \quad \rightarrow \text{hamiltonian-function } H(\mathbf{p},\mathbf{q}) = T(\mathbf{p})\;,$$

$$\mathbf{\Phi}_U^t \quad \leftrightarrow \quad \begin{cases} \dot{\mathbf{p}} &= -\,\mathbf{grad}\,U(\mathbf{q})\;, \\ \dot{\mathbf{q}} &= 0\;. \end{cases} \quad \rightarrow \text{hamiltonian-function } H(\mathbf{p},\mathbf{q}) = U(\mathbf{q})\;.$$

Corollary 4.4.13 $\Rightarrow$ $\mathbf{\Psi}^h$ is symplectic map ($\rightarrow$ Def. 4.4.12).

terminology: implicit midpoint-rule/Störmer-Verlet-method = symplectic integrators

$\triangle$

P. Grohs

rev 63606, February 17, 2014

Theorem 4.4.23 (symplectical Runge-Kutta-one-step method)**.**
All Runge-Kutta-one-step-methods ($\rightarrow$ Def. 2.3.5) which conserve quadratic invariants are symplectic.

Proof. $\mathbf{\Phi}^t \hat{=}$ Evolutionoperator are hamiltonian differential equation $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) := \mathbf{J}^{-1}\,\mathbf{grad}\,H(\mathbf{y})$ is a symplectic map for (all admissible) t

$I(\mathbf{Y}) := \mathbf{Y}^T\mathbf{J}\mathbf{Y}$ ist quadratic first integral of the variational equation

$$\dot{\mathbf{W}}(t;\mathbf{y}) = D\mathbf{f}(\mathbf{\Phi}^t\mathbf{y})\mathbf{W}(t;\mathbf{y})\ .$$

$\mathbf{\Psi}^h \hat{=}$ discrete evolution of the EK-OSM for $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$

$\widehat{\mathbf{\Psi}}^h \hat{=}$ discrete evolution of the EK-OSM for $\left\{\begin{array}{rcl}\dot{\mathbf{y}} &=& \mathbf{f}(\mathbf{y})\ ,\\ \dot{\mathbf{W}} &=& D\mathbf{f}(\mathbf{y})\mathbf{W}\end{array}\right.$: $\blacktriangleright$ $\left(\begin{smallmatrix}\mathbf{y}_1\\ \mathbf{W}_1\end{smallmatrix}\right) = \widehat{\mathbf{\Psi}}^h\left(\begin{smallmatrix}\mathbf{y}_0\\ \mathbf{I}\end{smallmatrix}\right)$.

Lemma 4.2.8 $\Rightarrow$ $$\boxed{\frac{d\mathbf{\Psi}^h}{d\mathbf{y}}(\mathbf{y}_0) = \mathbf{W}_1 = \left(\widehat{\mathbf{\Psi}}^h\begin{pmatrix}\mathbf{y}_0\\ \mathbf{I}\end{pmatrix}\right)_{\mathbf{W}}}\ .$$

P. Grohs

rev 63606, February 17, 2014

By assumption $\widehat{\mathbf{\Psi}}^h$ does contain quadratic first integral,

$$\left(\frac{d\mathbf{\Psi}^h}{d\mathbf{y}}(\mathbf{y}_0)\right)^T \mathbf{J} \left(\frac{d\mathbf{\Psi}^h}{d\mathbf{y}}(\mathbf{y}_0)\right) = \mathbf{W}_1^T \mathbf{J} \mathbf{W}_1 = \mathbf{J} \quad \forall \mathbf{y}_0 \in D \,. \qquad \square$$

Thm. 4.1.4 $\Rightarrow$ All Gauss-Collocations-one-step-method are symplectic.

Example 4.4.24 (Symplektisches Euler-Verfahren). siehe Bsp. 2.5.10

Separated Hamiltonian-function of the form $H(\mathbf{p},\mathbf{q}) = T(\mathbf{p}) + U(\mathbf{q}), T, U : D \subset \mathbb{R}^n \mapsto \mathbb{R}$ smooth

P. Grohs

rev 63606, February 17, 2014

$H(\mathbf{p},\mathbf{q}) = T(\mathbf{p}) + U(\mathbf{q})$ $\leftrightarrow$ Splitting of the right-hand side of (1.2.21), cf. Ex. 2.5.10

$$\mathbf{f}(\mathbf{y}) = \mathbf{J}^{-1}\,\mathbf{grad}\,H(\mathbf{y}) = \begin{pmatrix} -\,\mathbf{grad}\,U(\mathbf{q}) \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \mathbf{grad}\,T(\mathbf{p}) \end{pmatrix} =: \mathbf{f}_1(\mathbf{y}) + \mathbf{f}_2(\mathbf{y}) \tag{4.4.25}$$

➢ Lie-Trotter-Splitting-one-step method (2.5.2)

$$\begin{aligned} \mathbf{p}_{k+1} &= \mathbf{p}_k - h\,\mathbf{grad}\,U(\mathbf{q}_k) \\ \mathbf{q}_{k+1} &= \mathbf{q}_k + h\,\mathbf{grad}\,T(\mathbf{p}_{k+1}), \end{aligned} \quad \text{bzw.} \quad \begin{aligned} \mathbf{p}_{k+1} &= \mathbf{p}_k - h\,\mathbf{grad}\,U(\mathbf{q}_{k+1}) \\ \mathbf{q}_{k+1} &= \mathbf{q}_k + h\,\mathbf{grad}\,T(\mathbf{p}_k)\,. \end{aligned} \tag{4.4.26}$$

(4.4.26) = explicite symplectic discrete evolutions (Thm. 2.5.5: consistencyorder 1)

$\mathbf{q}_k$, $\mathbf{p}_{k+1}$

In (4.4.26) (right): Increment uses $\mathbf{q}_{k+1}$, $\mathbf{p}_k$ Generalization of (4.4.26) on $\dot{\mathbf{y}} = \mathbf{J}^{-1}\,\mathbf{grad}\,H(\mathbf{y})$ in the form, $\mathbf{y} := \binom{\mathbf{p}}{\mathbf{q}}$,

$$\dot{\mathbf{p}}(t) = -\frac{\partial H}{\partial \mathbf{q}}(\mathbf{p}(t),\mathbf{q}(t)) \quad , \quad \dot{\mathbf{q}}(t) = \frac{\partial H}{\partial \mathbf{p}}(\mathbf{p}(t),\mathbf{q}(t)) : \tag{1.2.21}$$

$$\mathbf{y}_{k+1} = \mathbf{y}_k + h\mathbf{J}^{-1}\,\mathbf{grad}\,H(\mathbf{p}_k,\mathbf{q}_{k+1}) \quad \text{bzw.} \quad \mathbf{y}_{k+1} = \mathbf{y}_k + h\mathbf{J}^{-1}\,\mathbf{grad}\,H(\mathbf{p}_{k+1},\mathbf{q}_k)\ .$$

P. Grohs

rev 63606, February 17, 2014

For general Hamiltonian-function $H = H(\mathbf{p},\mathbf{q})$: symplectic euler-method

$$\begin{aligned} \mathbf{p}_{k+1} &= \mathbf{p}_k - h\frac{\partial H}{\partial \mathbf{q}}(\mathbf{p}_{k+1},\mathbf{q}_k) \\ \mathbf{q}_{k+1} &= \mathbf{q}_k + h\frac{\partial H}{\partial \mathbf{p}}H(\mathbf{p}_{k+1},\mathbf{q}_k), \end{aligned} \quad \text{bzw.} \quad \begin{aligned} \mathbf{p}_{k+1} &= \mathbf{p}_k - h\frac{\partial H}{\partial \mathbf{q}}(\mathbf{p}_k,\mathbf{q}_{k+1}) \\ \mathbf{q}_{k+1} &= \mathbf{q}_k + h\frac{\partial H}{\partial \mathbf{p}}H(\mathbf{p}_k,\mathbf{q}_{k+1})\ . \end{aligned} \tag{4.4.27}$$

☞ no explicit splitting-method, despite symplecticity [21, Thm. 3.3]!

◇

Remark 4.4.28 (Partitionierte Runge-Kutta-Einschrittverfahren).

With local lipschitz-continuous $\mathbf{f}_u : D_u \times D_v \mapsto \mathbb{R}^n$, $\mathbf{f}_v : D_u \times D_v \mapsto \mathbb{R}^n$, $D_u, D_v \subset \mathbb{R}^n$

$$\text{ODE:} \quad \begin{aligned} \dot{\mathbf{u}} &= \mathbf{f}_u(\mathbf{u}, \mathbf{v}) \ , \\ \dot{\mathbf{v}} &= \mathbf{f}_v(\mathbf{u}, \mathbf{v}) \ . \end{aligned} \tag{4.4.29}$$

"symplectic Euler-method" (4.4.27) for (4.4.29):

$$\begin{aligned} \mathbf{u}_1 &= \mathbf{u}_0 + h\mathbf{f}_u(\mathbf{u}_1, \mathbf{v}_0) \ , \\ \mathbf{v}_1 &= \mathbf{v}_0 + h\mathbf{f}_v(\mathbf{u}_1, \mathbf{v}_0) \ . \end{aligned} \quad \succ \quad \text{Consistencyorder 1.} \tag{4.4.30}$$

Ansatz: s-level partitioned Runge-Kutta-one-step method (for autonomous ODE)

$$\begin{cases} \mathbf{k}_i^u &= \mathbf{f}_u(\mathbf{u}_0 + h \sum\limits_{j=1}^{s} a_{ij}^u \mathbf{k}_j^u, \mathbf{v}_0 + h \sum\limits_{j=1}^{s} a_{ij}^v \mathbf{k}_j^v) \ , \\ \mathbf{k}_i^v &= \mathbf{f}_v(\mathbf{u}_0 + h \sum\limits_{j=1}^{s} a_{ij}^u \mathbf{k}_j^u, \mathbf{v}_0 + h \sum\limits_{j=1}^{s} a_{ij}^v \mathbf{k}_j^v) \end{cases} \qquad i = 1, \dots, s \ ,$$

$$\begin{cases} \mathbf{u}_1 &= \mathbf{u}_0 + \sum\limits_{i=1}^{s} b_i^u \mathbf{k}_i^u \ , \\ \mathbf{v}_1 &= \mathbf{v}_0 + \sum\limits_{i=1}^{s} b_i^v \mathbf{k}_i^v \ . \end{cases} \tag{4.4.31}$$

in echolon form, cf. Rem. 2.3.7:

$$\left\{\begin{array}{rcl} \mathbf{g}_i^u &=& \mathbf{u}_0 + h\sum_{j=1}^{s} a_{ij}^u \mathbf{f}_u(\mathbf{g}_j^u,\mathbf{g}_j^v)\;, \\ \mathbf{g}_i^v &=& \mathbf{v}_0 + h\sum_{j=1}^{s} a_{ij}^v \mathbf{f}_v(\mathbf{g}_j^u,\mathbf{g}_j^v)\;, \end{array}\right. \quad , \quad \left\{\begin{array}{rcl} \mathbf{u}_1 &=& \mathbf{u}_0 + \sum_{i=1}^{s} b_i^u \mathbf{f}_u(\mathbf{g}_i^u,\mathbf{g}_i^v)\;, \\ \mathbf{v}_1 &=& \mathbf{v}_0 + \sum_{i=1}^{s} b_i^v \mathbf{f}_v(\mathbf{g}_i^u,\mathbf{g}_i^v)\;. \end{array}\right. \tag{4.4.32}$$

Illustration: Two Butcher-Tableaus:

$$\begin{array}{c|c} \mathbf{c}^u & \mathfrak{A}^u \\ \hline & \mathbf{b}^{u,T} \end{array} \quad \& \quad \begin{array}{c|c} \mathbf{c}^v & \mathfrak{A}^v \\ \hline & \mathbf{b}^{v,T} \end{array}$$

Symplectic Euler-method

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \quad \& \quad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

Störmer-Verlet-method

$$\begin{array}{c|cc} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 1/2 & 1/2 \end{array} \quad \& \quad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}$$

In Analogy to the theory of conventional RK-OSM from Def. 2.3.5:

- Equations on the coefficients for desired consistency order, cf. Sect. 2.3.2 [21, Sect. II.2]
- algebraic Conditions for the Conservation of quadratic invariants [21, Sect. IV.2.2], cf. Lemma 4.1.6, and symmetry, cf. Thm. 4.3.3 [21, Sect. V.2.2],
- Conditions on the coefficients for symplecticity, cf. Thm. 4.4.23 [21, Sect. VI.4].

Example 4.4.33 (Symplectic Euler-method for pendulum equation).

IVP for pendulum equation as in Ex. 4.4.3.

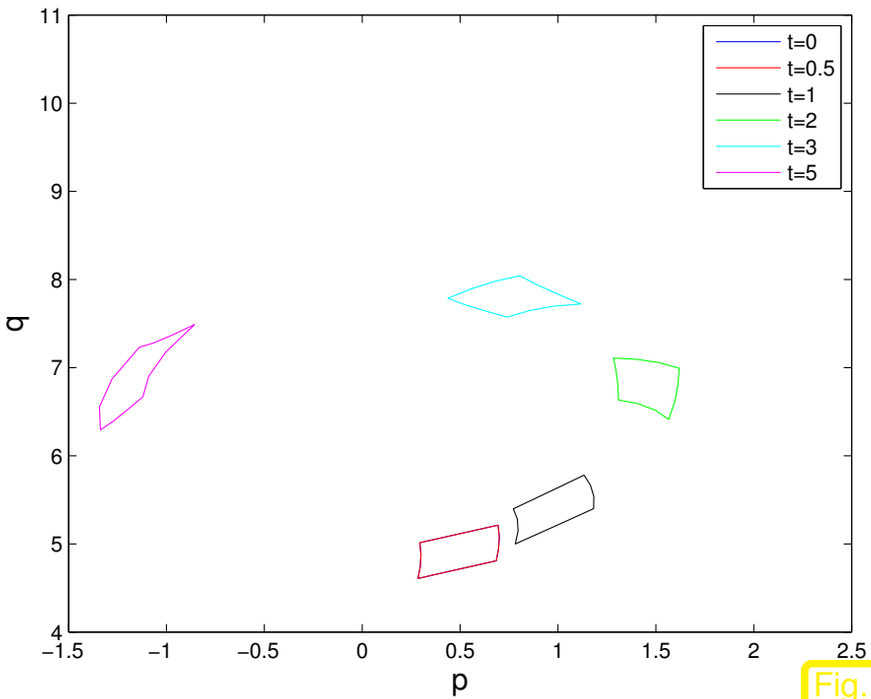


Fig. 164

Evolution of a quadratic volume (Method (4.4.27), left)

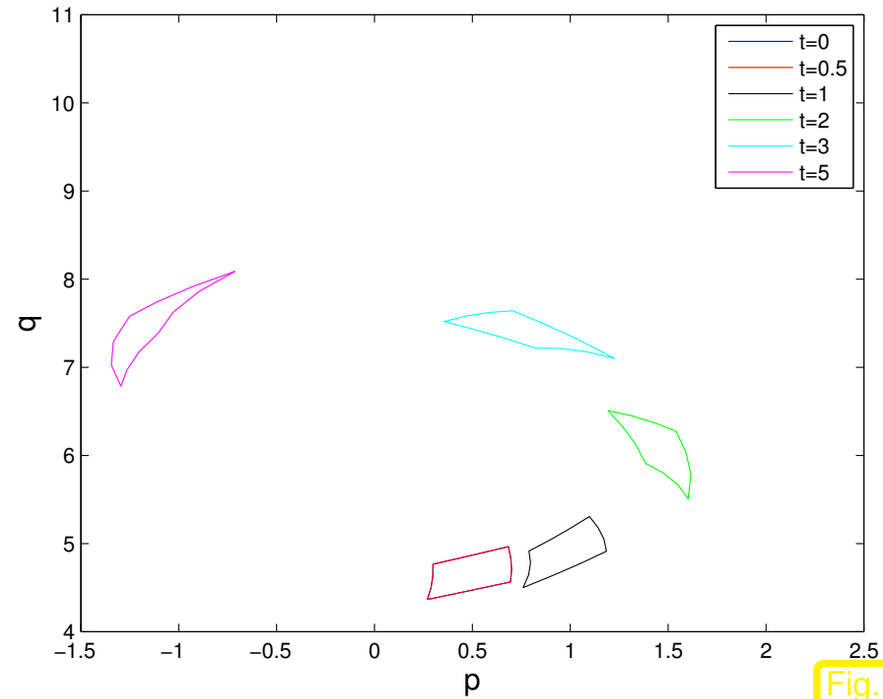


Fig. 165

Evolution of a quadratic volume (Method (4.4.27), right)

Conservation of energy of the symplectical partitioned euler-method (4.4.27) (left) ($p(0) = 0, q(0) = \frac{7\pi}{6}$)

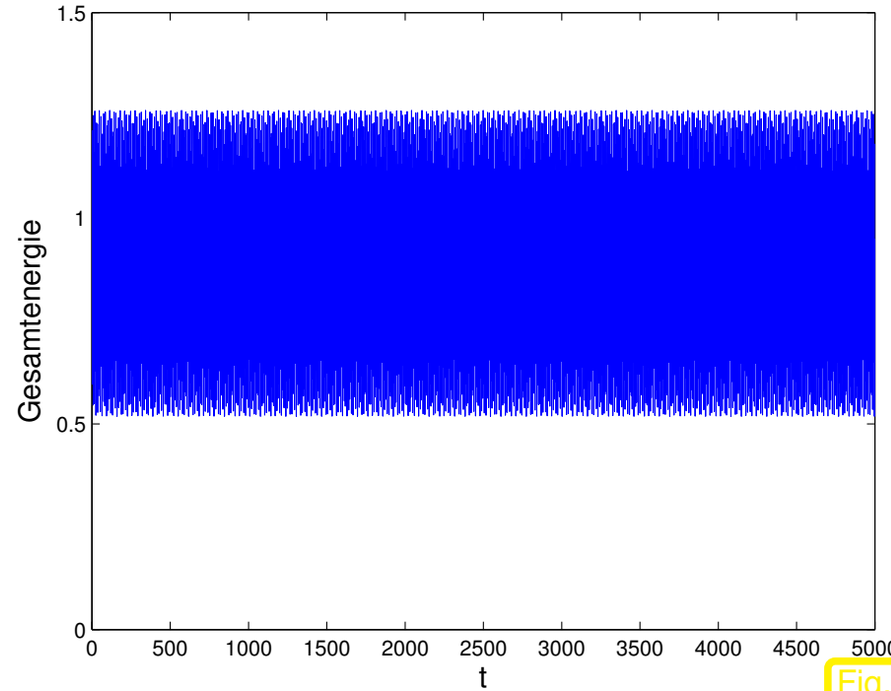


Fig. 166

◇

Example 4.4.34 (Longtime-energy conservation with symplectical integration). → Ex. 4.4.1, 4.4.33, 1.4.32

Hamiltonian differential equation (4.4.4) for mathematical pendulum → 1.2.17 ($p \leftrightarrow$ angular velocity, $q \leftrightarrow$ angular variable α)

$$\begin{aligned} \dot{p} &= -\sin q \,, \\ \dot{q} &= p \end{aligned} \qquad \blacktriangleright \quad \text{Hamiltonian-function } H(p,q) = \frac{1}{2}p^2 - \cos q \quad \text{(Total energy)} \qquad (4.4.4)$$

Initial values: $p(0) = 0$, $q(0) = 7/6\pi$, Endtime $T = 5000$

- Symplectic partitioned Euler-method (4.4.27) (left)
- Störmer-Verlet-method (4.4.22), see Rem. 4.4.19
- Implicit midpoint rule (4.4.20), see Rem. 4.4.19
- 2-level Gauss-collocations-one-step method, see Sect. 2.2.1

(uniform time-stepwidth $h > 0$)

strength of change in energy

$$E_{\text{var}}(h) = \max_{i=0,\dots,T/h} |E_h(ih) - E_{\text{exact}}| \ .$$

Sympl. Euler	$E_{\text{var}}(h) = O(h)$,
Störmer-Verlet	$E_{\text{var}}(h) = O(h^2)$,
Implizite MPR	$E_{\text{var}}(h) = O(h^2)$,
Gauss-Koll ($s = 2$)	$E_{\text{var}}(h) = O(h^4)$.

$E_{\text{var}}(h) = O(h^p)$

($p \hat{=}$ Convergence order of the OSM)

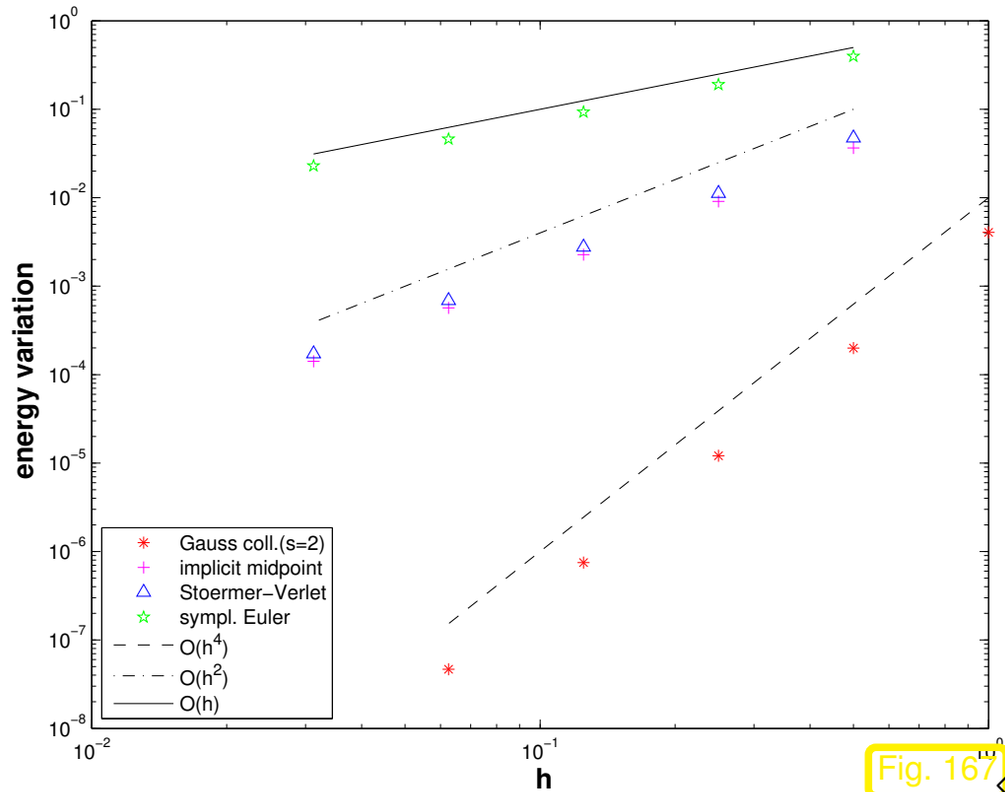


Fig. 167 ◇

Example 4.4.35 (Spring pendulum).

Frictionless spring pendulum: hamiltonian-function $H(\mathbf{p},\mathbf{q}) = \frac{1}{2}\|\mathbf{p}\|^2 + \frac{1}{2}(\|\mathbf{q}\| - 1)^2 + q_2$

($\mathbf{q} \hat{=}$ position, $\mathbf{p} \hat{=}$ impulse)

$$\dot{\mathbf{p}} = -(\|\mathbf{q}\| - 1)\frac{\mathbf{q}}{\|\mathbf{q}\|} - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad , \quad \dot{\mathbf{q}} = \mathbf{p} \; . \tag{4.4.36}$$

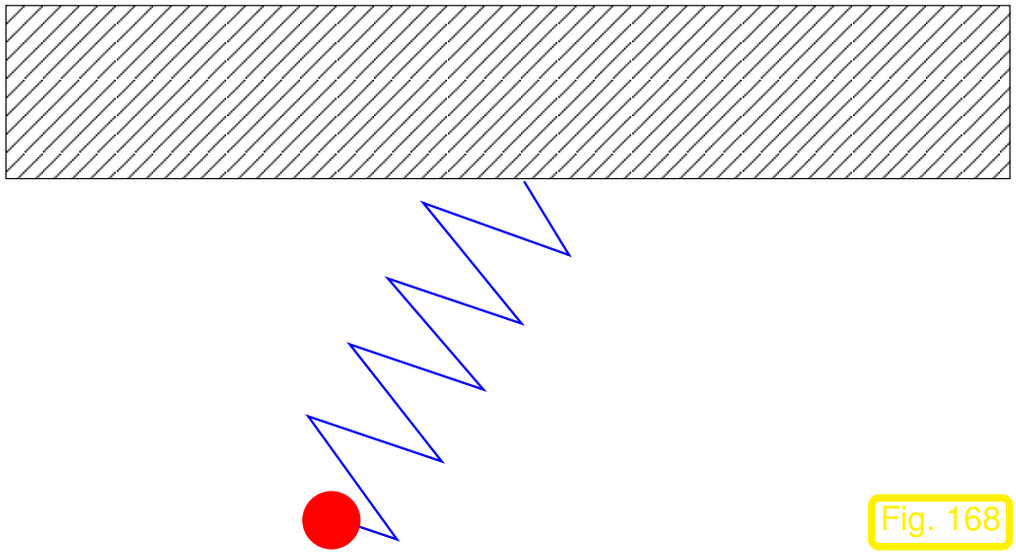


Fig. 168

Trajectories of longtime evolution ▷

(Chaotic mechanical system)

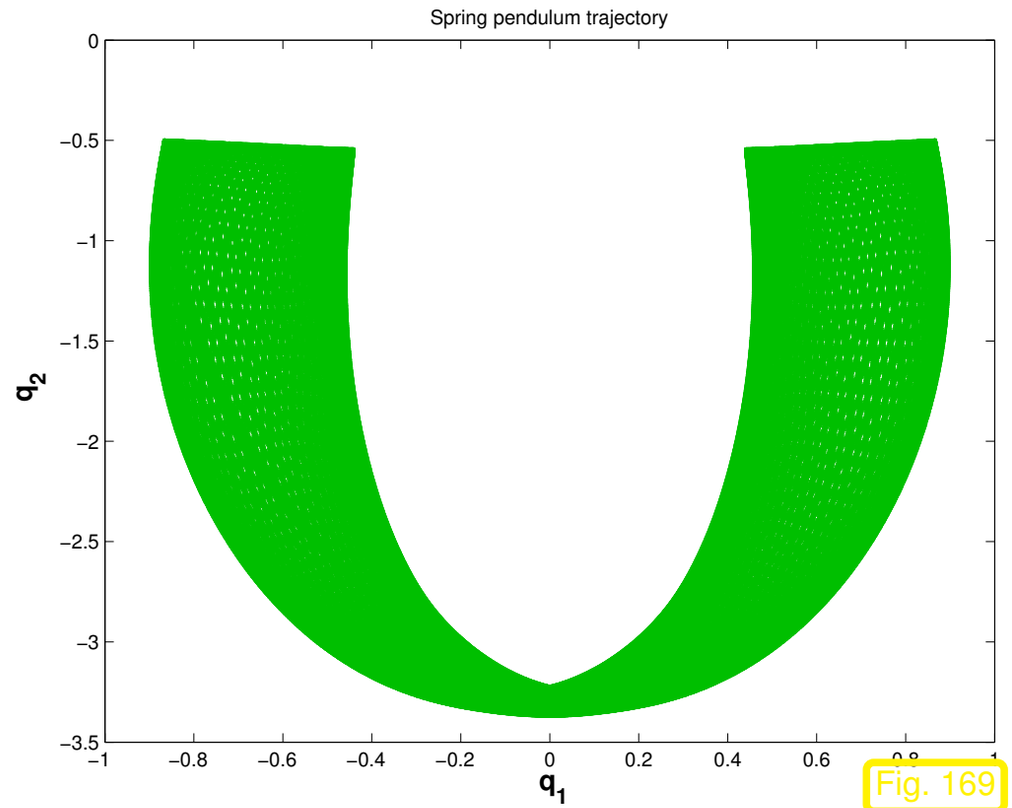
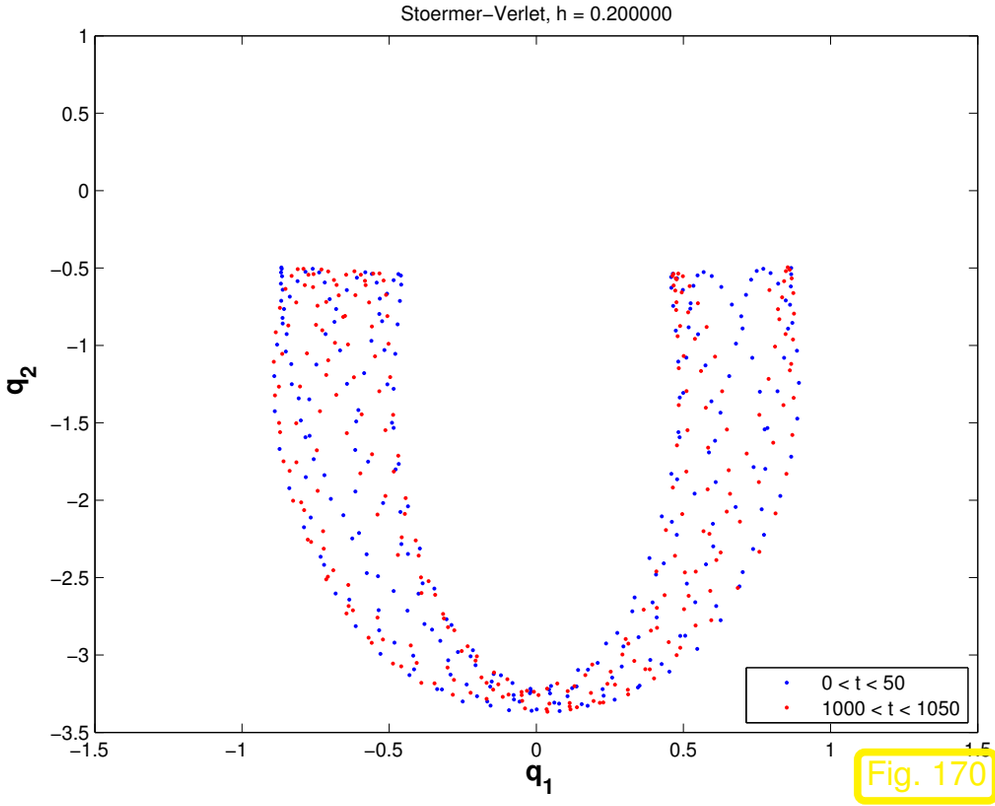
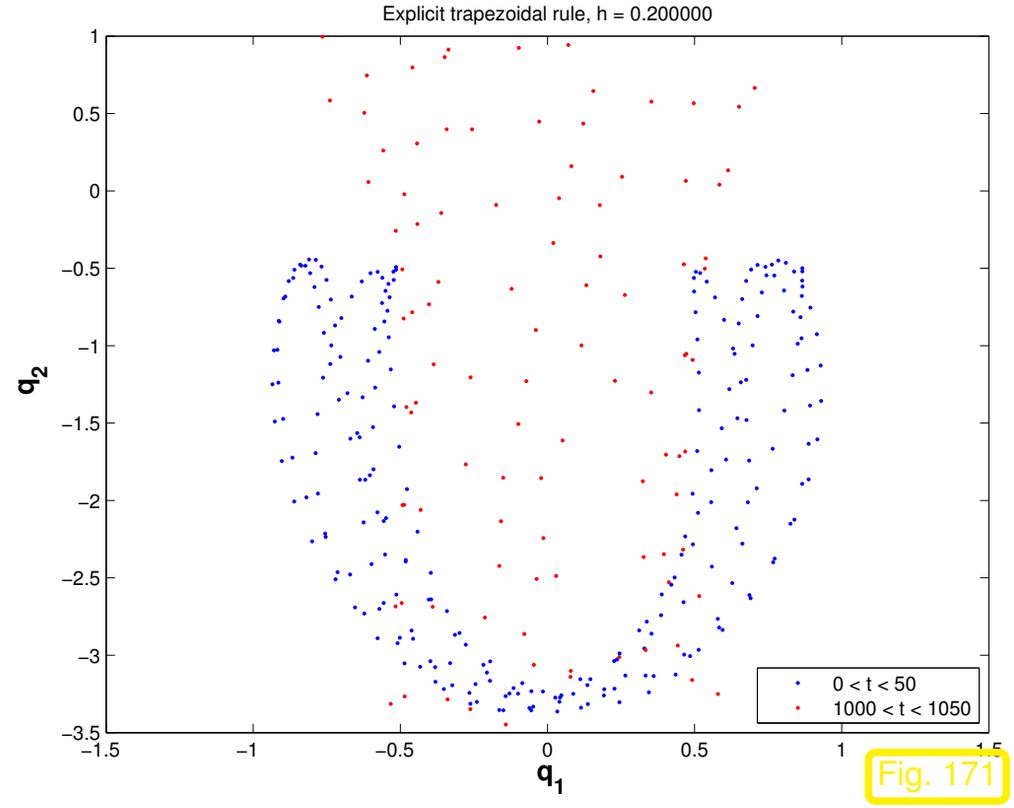


Fig. 169

- Störmer-Verlet-method (4.4.22) (Consistenceorder 2), see Rem. 4.4.19,
- Explicit trapezoidal rule (2.3.3) (Consistenceorder 2).



Störmer-Verlet OSM



Explicit trapezoidal rule

Animation ▷

symplectic integrator: position in a “admissible region” also in a longtime-integration

Explicit trapezoidal rule: trajectories leave the longtime integration “admissible region” (energy-shift !)

Example 4.4.37 (Molecular dynamic). $\rightarrow$ [12, Sect. 1.2]

Phase space for $n \in \mathbb{N}$ atoms in $d \in \mathbb{N}$ dimensions: $D = \mathbb{R}^{2dn}$
(positions $\mathbf{q} = [\mathbf{q}^1; \dots ; \mathbf{q}^n]^T \in \mathbb{R}^{dn}$, impulse $\mathbf{p} = [\mathbf{p}^1, \dots , \mathbf{p}^n]^T \in \mathbb{R}^{dn}$)

Total energy (Hamiltonian-function):

$$H(\mathbf{p}, \mathbf{q}) = \frac{1}{2} \|\mathbf{p}\|_2^2 + V(\mathbf{q}) \ .$$

Lenard-Jones-potential: ▷

$$V(\mathbf{q}) = \sum_{j=1}^{n} \sum_{i \neq j} \mathcal{V}(\left\| \mathbf{q}^i - \mathbf{q}^j \right\|_2) \ ,$$

$$\mathcal{V}(\xi) = \xi^{-12} - \xi^{-6} \ . \qquad (4.4.38)$$

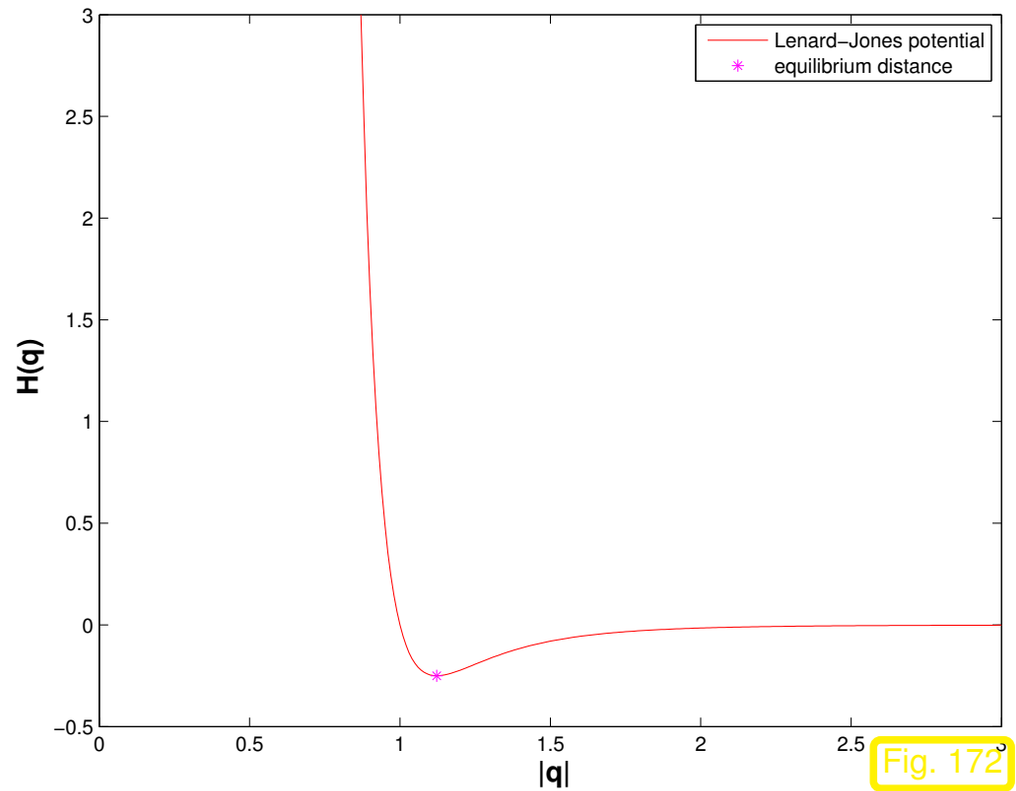


Fig. 172

➥ hamiltonian differential equation (→ def. 1.2.20):

$$\dot{\mathbf{p}}^j = -\sum_{i\neq j} \mathcal{V}'(\left\|\mathbf{q}^j - \mathbf{q}^i\right\|_2)\frac{\mathbf{q}^j - \mathbf{q}^i}{\left\|\mathbf{q}^j - \mathbf{q}^i\right\|_2} \quad , \quad \dot{\mathbf{q}}^j = \mathbf{p}^j \ , j = 1, \dots, n \ .$$

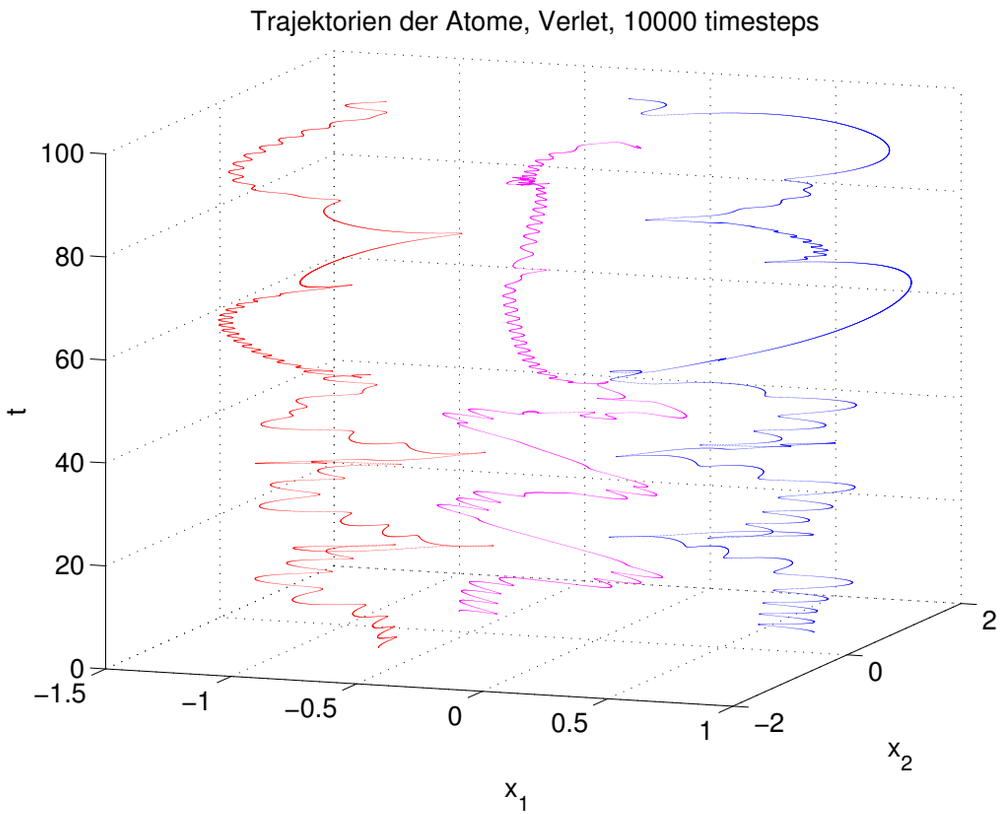
▶ Störmer-Verlet-method (4.4.22):

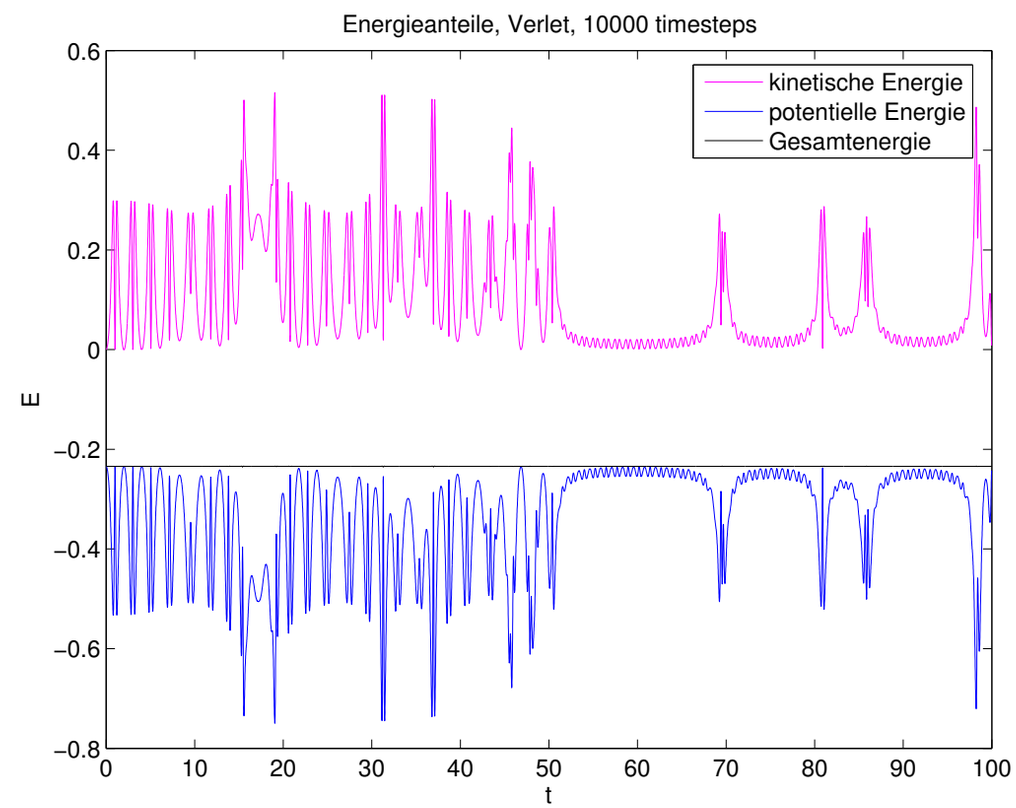
$$\mathbf{q}_h(t+\tfrac{1}{2}h) = \mathbf{q}_h(t) + \tfrac{h}{2}\mathbf{p}_h(t) \ ,$$

$$\mathbf{p}_h^j(t+h) = \mathbf{p}_h^j(t) - h\sum_{i\neq j} \mathcal{V}'(\left\|\mathbf{q}_h^j(t+\tfrac{1}{2}h) - \mathbf{q}_h^i(t+\tfrac{1}{2}h)\right\|_2)\frac{\mathbf{q}_h^j(t+\frac{1}{2}h) - \mathbf{q}_h^i(t+\frac{1}{2}h)}{\left\|\mathbf{q}_h^j(t+\frac{1}{2}h) - \mathbf{q}_h^i(t+\frac{1}{2}h)\right\|_2} \ ,$$

$$\mathbf{q}_h(t+h) = \mathbf{q}_h(t+\tfrac{1}{2}h) + \tfrac{h}{2}\mathbf{p}_h(t+h) \ .$$

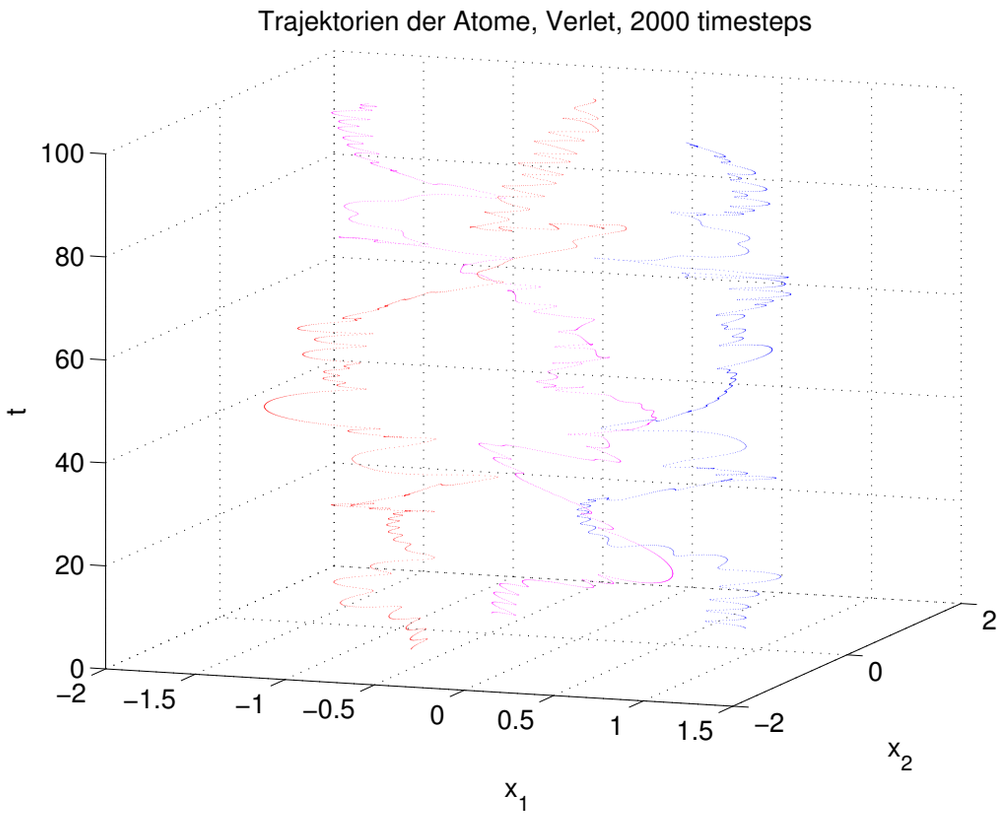
Simulation with $d = 2$, $n = 3$, $\mathbf{q}^1(0) = \frac{1}{2}\sqrt{2}\binom{-1}{-1}$, $\mathbf{q}^2(0) = \frac{1}{2}\sqrt{2}\binom{1}{1}$, $\mathbf{q}^3(0) = \frac{1}{2}\sqrt{2}\binom{-1}{1}$, $\mathbf{p}(0) = 0$, endtime $T = 100$

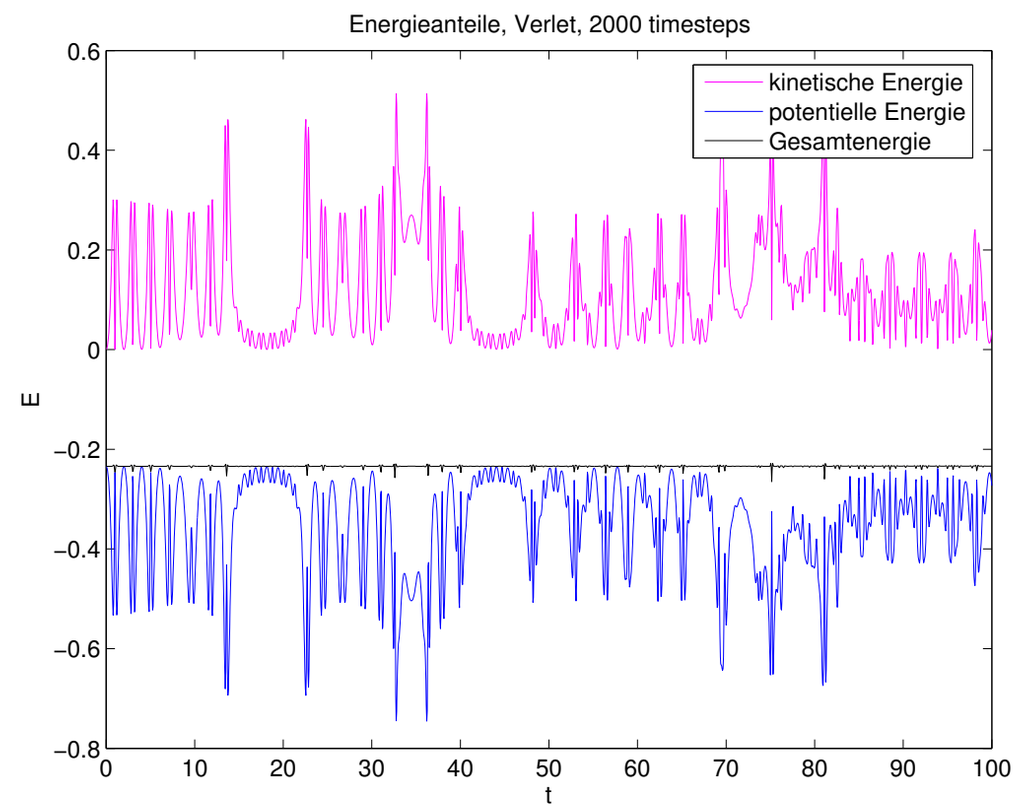




P. Grohs

rev 63606, February 17, 2014





P. Grohs

rev 63606, February 17, 2014

Observation:

- Completely different trajectories of longtime simulation with different time stepwidths h.
- Qualitatively correct trajectories in each case.

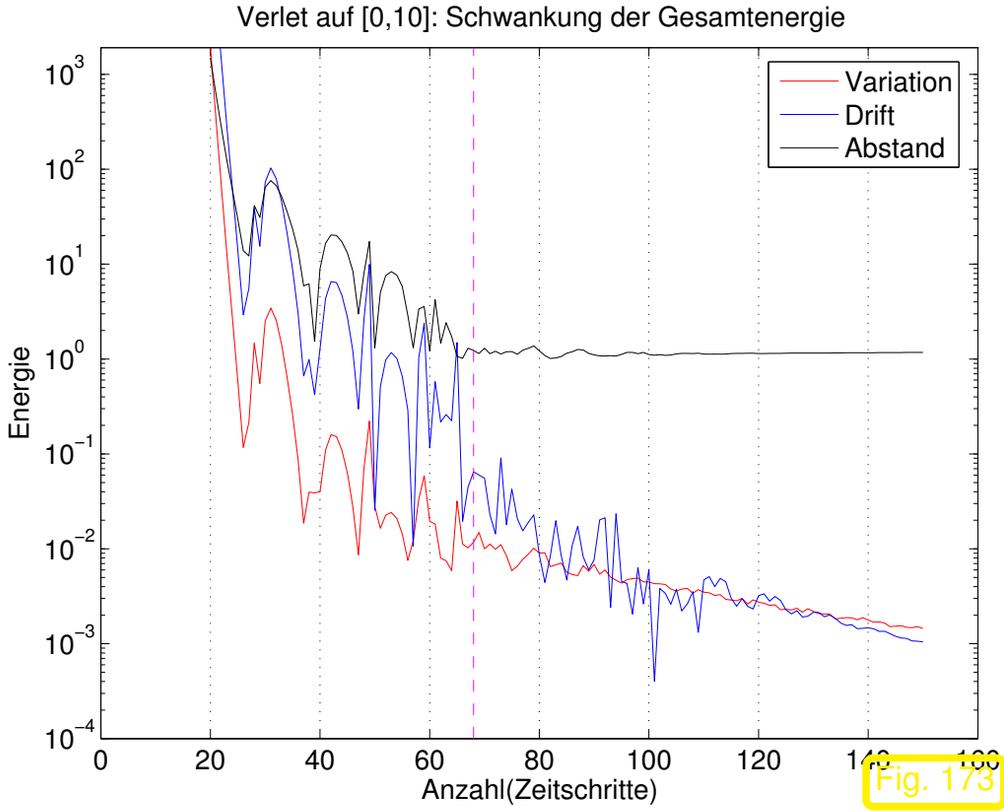


Fig. 173

$T = 10$, $d = 2$, $n = 3$, $\mathbf{q}^1(0) = \frac{1}{2}\sqrt{2}\binom{-1}{-1}$, $\mathbf{q}^2(0) = \frac{1}{2}\sqrt{2}\binom{1}{1}$, $\mathbf{q}^3(0) = \frac{1}{2}\sqrt{2}\binom{-1}{1}$, $\mathbf{p}(0) = 0$.

$$\text{Variation} = \sum_{i=1}^{N-1} |E_{\text{tot}}((i+1)h) - E_{\text{tot}}(ih)| \,,$$

$$\text{Drift} = |E_{\text{tot}}(T) - E_{\text{tot}}(0)| \,,$$

$$\text{Distance} = \max\{\left\|\mathbf{q}_h^j(T)\right\|_2, j = 1,2,3\} \,.$$

◇

Example 4.4.39 (Multipleparticle-molecular dynamics). → [31, Sect. 4.5.1]

2D conservative Multipleparticle-system with Lennard-Jones-Potential $\rightarrow$ Ex. 4.4.37

Initial positions $\triangleright$

(Initial impulses $= 0 \leftrightarrow$ 0K)

Observed for explicit trapezoidal rule (2.3.3), Störmer-Verlet (4.4.22)

- Approximation of total energy $H(\mathbf{p}, \mathbf{q})$
- mean kinetic energy ("temperature")

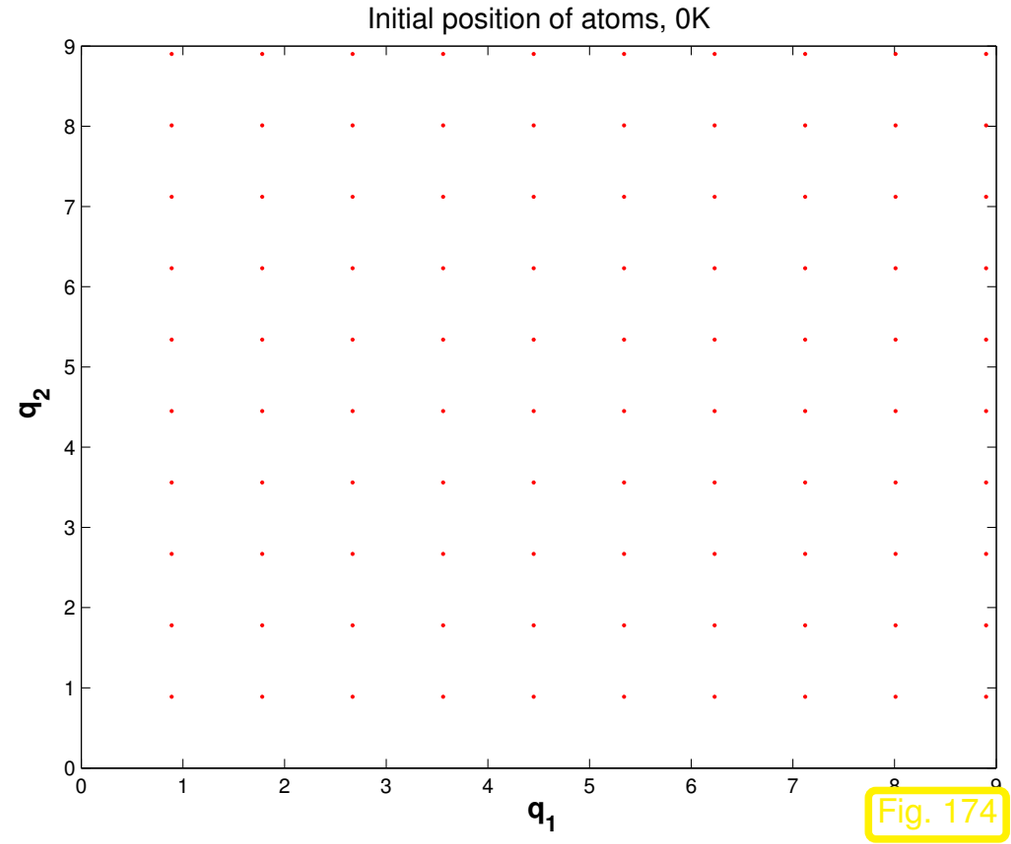


Fig. 174

Animation $\triangleright$

P. Grohs

rev 63606, February 17, 2014

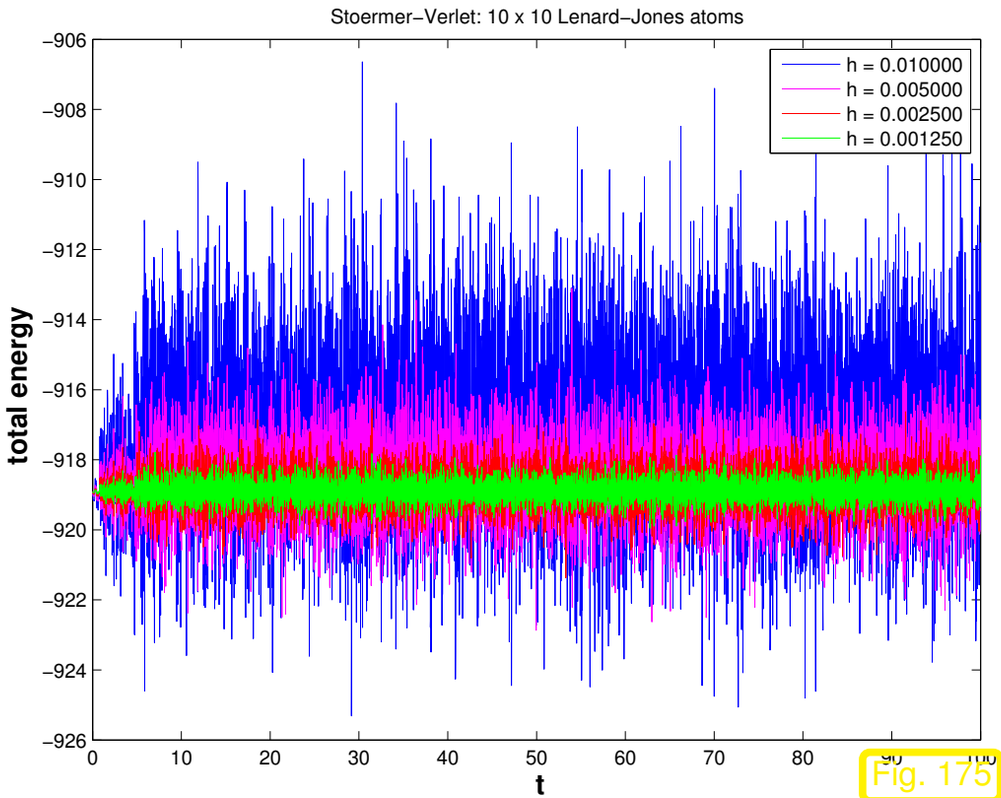


Fig. 175

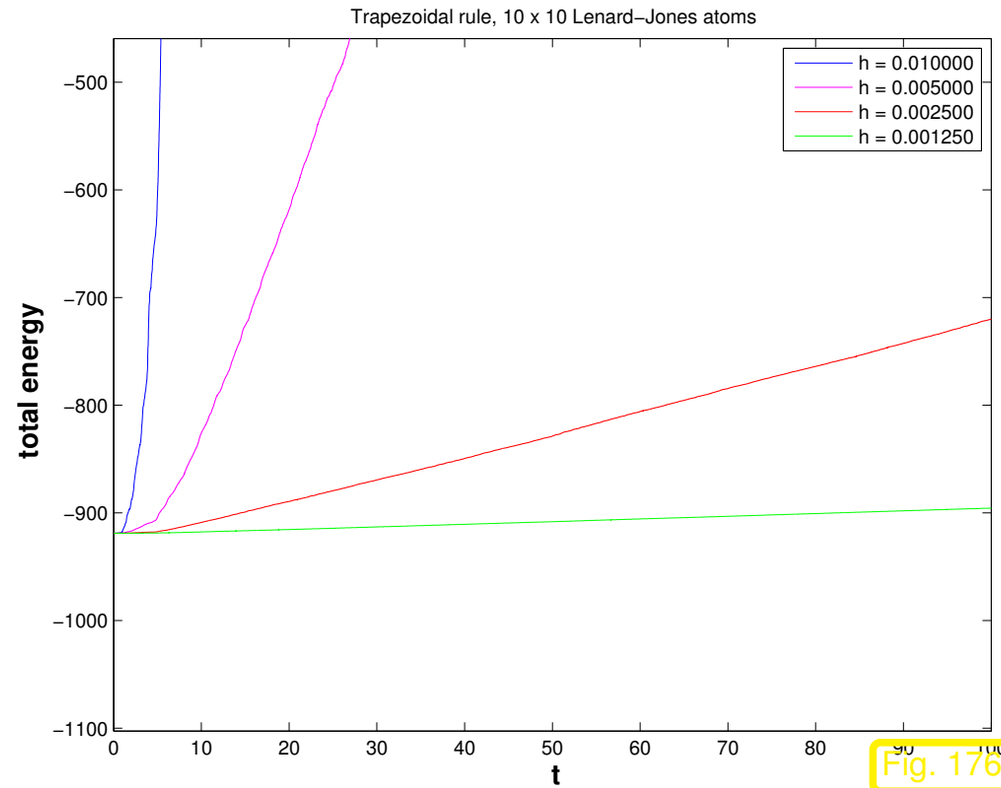


Fig. 176

P. Grohs

rev 63606, February 17, 2014

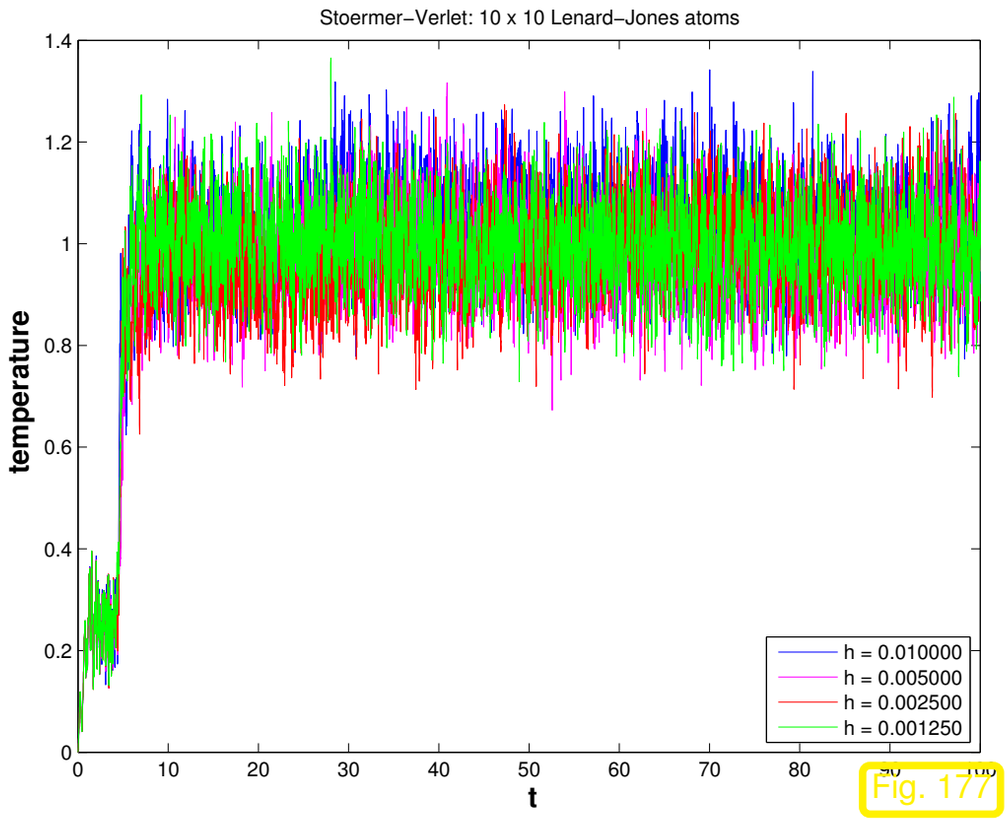


Fig. 177

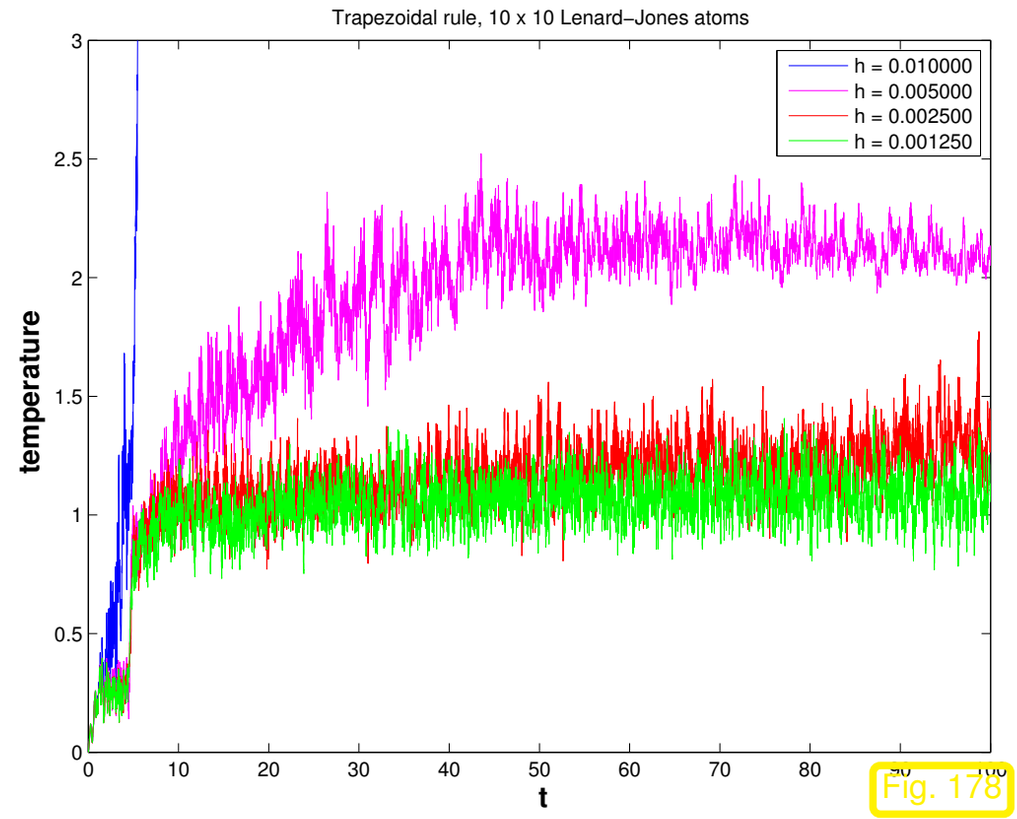


Fig. 178

P. Grohs

rev 63606, February 17, 2014

Symplectic integrator: qualitatively correct behavior of temperature

◇

Example 4.4.40 (projection on manifold of energy). → Ex. 4.4.35

Correction of energy-shift (with non-symplectical integrators) by *projection* on manifold of energy

$$\{(\mathbf{p},\mathbf{q}) \in \mathbb{R}^n \times \mathbb{R}^n \colon H(\mathbf{p},\mathbf{q}) = H(\mathbf{p}_0,\mathbf{q}_0)\} \; . \tag{4.4.41}$$

More concrete: Orthogonal projection $(\mathbf{p},\mathbf{q}) \mapsto \mathsf{P}(\mathbf{p},\mathbf{q}) := (\mathbf{p}^*,\mathbf{q}^*)$: with $\mathbf{y} = \binom{\mathbf{p}}{\mathbf{q}}$, determine $\lambda \in \mathbb{R}$, $\mathbf{y}^* = \binom{\mathbf{p}^*}{\mathbf{q}^*} \in \mathbb{R}^{2n}$ such that

$$H(\mathbf{y}^*) = H_0 \quad , \quad \mathbf{y}^* = \mathbf{y} + \lambda \, \mathbf{grad}\, H(\mathbf{y}^*) \; . \tag{4.4.42}$$

projected OSM $\mathbf{\Psi}^h$: orthogonal projection after each step: $\mathbf{y}_{k+1} = \mathsf{P}\mathbf{\Psi}^h \mathbf{y}_k$

Note: (4.4.42) nonlinear system of equation of dimension $2n+1$, expensive !

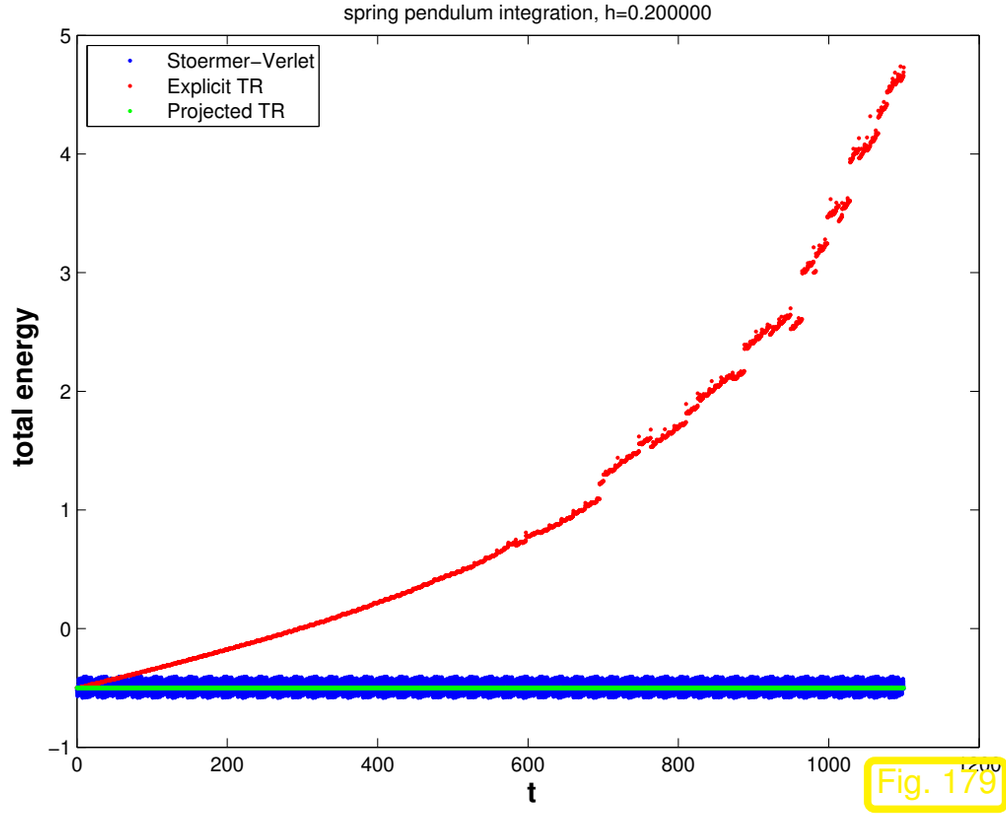


Fig. 179

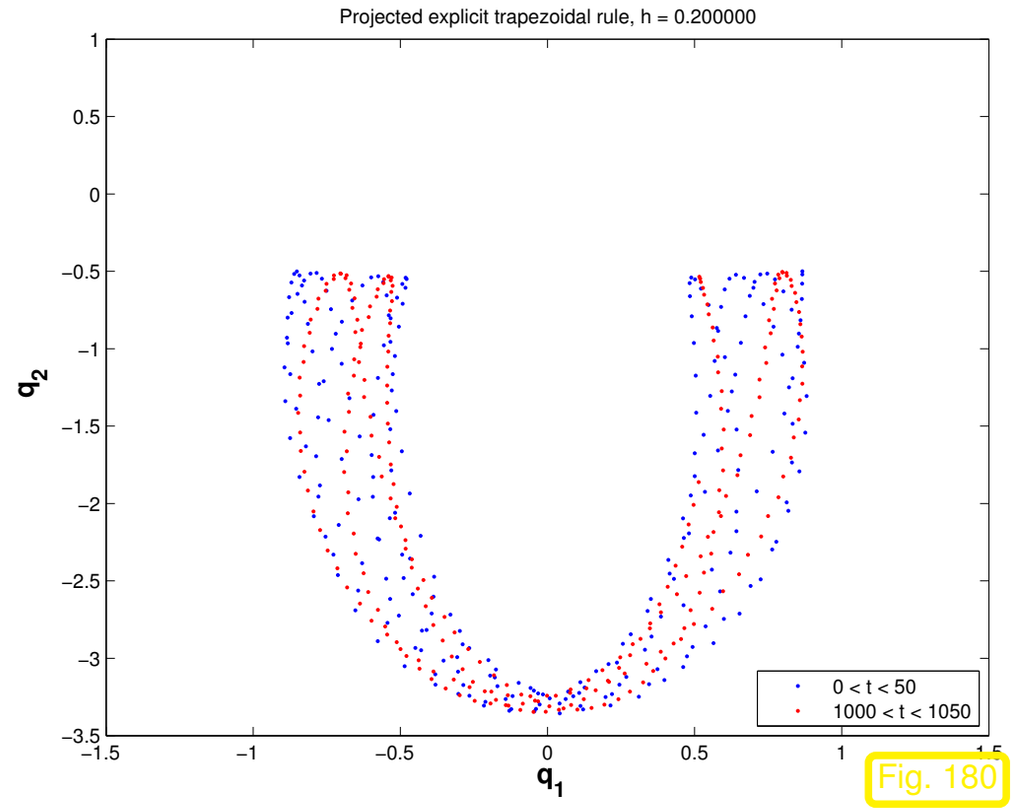


Fig. 180

P. Grohs

rev 63606, February 17, 2014

Attention: projection not a universal remedy, see [21, Ch. IV, Ex. 4.3]

◇

4.4.3 Backward analysis

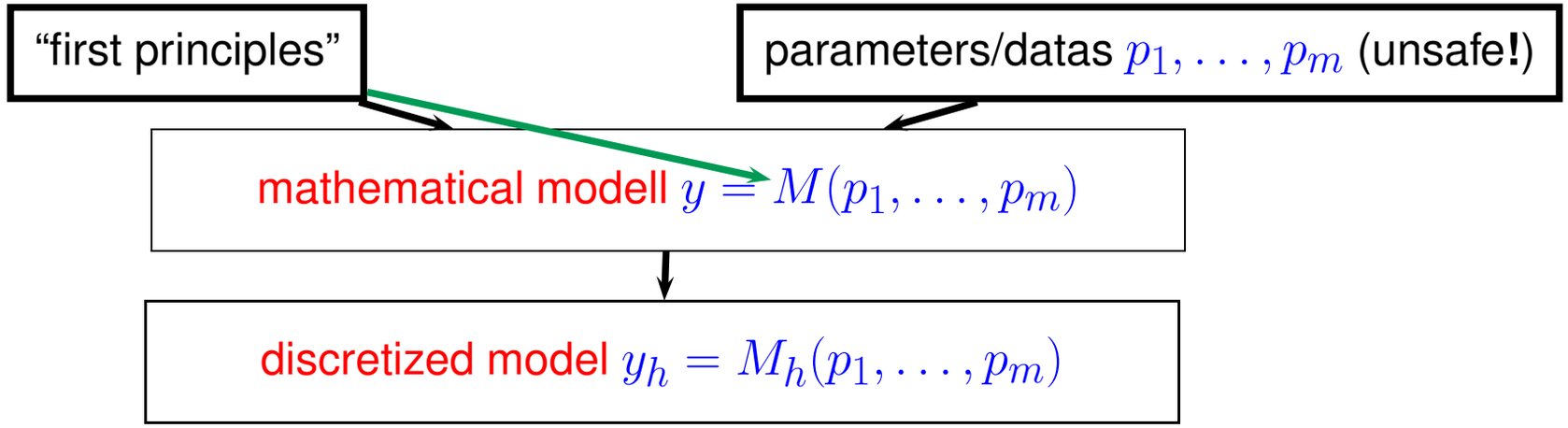
Sect. 1.3.3.5: Computation of individual trajectories useless for ill-conditioned/chaotic evolutions.

Goal: Computation of typical/chaotic evolution.

Computation of individual trajectories useless for ill-conditioned chaotic evolutions.

Goal: Computation of typical/probable trajectories

Remark 4.4.43 (Backward error analysis: philosophical basics).



Goal: Exact solution y_h which is consistent with the first principles

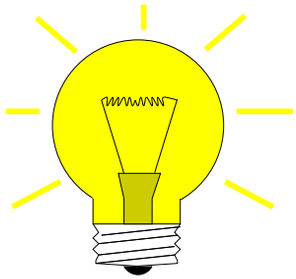
$$\exists \widetilde{p}_i: \quad \|p_i - \widetilde{p}_i\| \ll 1: \quad y_h = M(\widetilde{p}_1, \ldots, \widetilde{p}_m)$$

△

P. Grohs

rev 63606, February 17, 2014

Concrete application of this philosophy on numerical integrator (One-step-method), see [31, Sect. 5.1]:



$\mathbf{\Psi}^h \hat{=}$ discrete Evolution of a OSM for $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) \quad \succ \quad \mathbf{y}_{k+1} = \mathbf{\Psi}^h(\mathbf{y}_k)$

Find h-dependent vectorfield $\widetilde{\mathbf{f}}_h :\mapsto \mathbb{R}^d$ such that

$$\dot{\widetilde{\mathbf{y}}} = \widetilde{\mathbf{f}}_h(\widetilde{\mathbf{y}}) \;, \quad \widetilde{\mathbf{y}}(0) = \mathbf{y}_0 \quad \Rightarrow \quad \mathbf{y}_k = \widetilde{\mathbf{y}}(hk) \;. \tag{4.4.44}$$

modified differential equation

Wish:

- $\widetilde{\mathbf{f}}_h$, $\mathbf{f}$ belongs to the same class of vectorfields, cf. rem. 4.4.14
- Small perturbation: $\widetilde{\mathbf{f}}_h \approx \mathbf{f}$ for "small" stepwidths $h > 0$

$\mathbf{\Psi}^h$ structure-preserving & "qualitatively exact": $(\mathbf{y}_k)$ acceptable

Backward error analysis on the basis of modified differential equation requires uniform time-stepwidth

Example 4.4.45 (Modifizierte Gleichung for RK-ESV und lineare ODE).

- linear ODE ($\rightarrow$ Sect. 1.3.2): $\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}$, $\mathbf{A} \in \mathbb{R}^{d,d}$

- Runge-Kutta-one-step method ($\rightarrow$ Def. 2.3.5) with stability function $S(z)$

$$\blacktriangleright \quad \text{modified ODE:} \quad \widetilde{\mathbf{f}}_h(\mathbf{y}) = \widetilde{\mathbf{A}}\mathbf{y} \quad , \quad \widetilde{\mathbf{A}} = \frac{1}{h}\log(S(h\mathbf{A})) \; , \tag{4.4.46}$$

for "sufficiently small" $h > 0$.

Here: $\log \hat{=}$ "matrixlogarithm": $\log(\mathbf{X}) = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k}(\mathbf{X}-\mathbf{I})^k$ for $\|\mathbf{X}-\mathbf{I}\| < 1$

Proof of (4.4.46) (elementary under assumption that $\mathbf{A}$ diagonalisable: $\exists \mathbf{T} \in \mathbb{R}^{d,d}$ regular: $\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \mathbf{D} = \operatorname{diag}(\mu_1, \dots, \mu_d)$):

Rem. 3.1.13, (3.1.16) $\Rightarrow$ For RK-OSM $\mathbf{y}_1 = S(h\mathbf{A})\mathbf{y}_0$

$$\overset{(4.4.44)}{\Longrightarrow} \quad \exp(\widetilde{\mathbf{A}}h) = S(h\mathbf{A}) \quad \text{with} \quad \sigma(h\mathbf{A}) \cap]-\infty, 0] = \emptyset \quad \text{for small } h > 0 \; .$$

$\diamond$

? Modified equation in the general case

Definition 4.4.47 (Modified equation of order q)**.**
Let $\mathbf{\Psi}^h$ be the discrete evolution of a one-step method der order of consistency p for the ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ with local lipschitz-continuous $\mathbf{f} : D \subset \mathbb{R}^d \mapsto \mathbb{R}^d$.
Then $\dot{\widetilde{\mathbf{y}}} = \widetilde{\mathbf{f}}_h(\widetilde{\mathbf{y}})$ is together with h-dependent, local lipschitz-continuous $\widetilde{\mathbf{f}}_h : D \mapsto \mathbb{R}^d$ a modified equation of order q, $q > p$, if

$$\left\| \widetilde{\mathbf{\Phi}}_h^h \mathbf{y} - \mathbf{\Psi}^h \mathbf{y} \right\| \leq C(\mathbf{y}) h^{q+1} \quad \forall \mathbf{y} \in D \quad \text{for} \quad h \to 0 \ ,$$

where $\widetilde{\mathbf{\Phi}}_h^t$ is the evolution-operator to $\dot{\widetilde{\mathbf{y}}} = \widetilde{\mathbf{f}}_h(\widetilde{\mathbf{y}})$ and $C : D \mapsto \mathbb{R}$ locally uniformly bounded.

Def. 4.4.47 $\hat{=}$ "The OSM is consistent of order q with $\dot{\widetilde{\mathbf{y}}} = \widetilde{\mathbf{f}}_h(\widetilde{\mathbf{y}})$." $\to$ Def. 2.1.13

P. Grohs

rev 63606, February 17, 2014

Example 4.4.48 (Modified equation of order 2 to explicit euler-method).

explicite euler-method (1.4.2) for $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$: $\mathbf{y}_1 = \mathbf{y}_1(h) = \mathbf{\Psi}^h \mathbf{y}_0 = \mathbf{y}_0 + h\mathbf{f}(\mathbf{y}_0)$

Comparison with taylorexpansion (around 0) (2.3.25) of the exact solution $\mathbf{y}(t)$:

$$\mathbf{y}(h) = \mathbf{y}_0 + \mathbf{f}(\mathbf{y}_0)h + \tfrac{1}{2}D\mathbf{f}(\mathbf{y}_0)\mathbf{f}(\mathbf{y}_0)h^2 + O(h^3) \quad \text{for } h \to 0 \,. \tag{4.4.49}$$

"perturbed term ☞ to "offset" in $\widetilde{\mathbf{f}}_h$

Modified equation of order 2:

$$\dot{\widetilde{\mathbf{y}}} = \widetilde{\mathbf{f}}_h(\widetilde{\mathbf{y}}) := \mathbf{f}(\widetilde{\mathbf{y}}) - \tfrac{1}{2}hD\mathbf{f}(\widetilde{\mathbf{y}})\mathbf{f}(\widetilde{\mathbf{y}}) \quad \overset{\mathbf{f}\text{ glatt}}{\Longrightarrow} \quad \widetilde{\mathbf{\Phi}}_h^h \mathbf{y}_0 - \mathbf{y}_1(h) = O(h^3) \,,$$

then from (4.4.49), fpr $h \to 0$

$$\begin{aligned}
\widetilde{\mathbf{y}}(h) &= \mathbf{y}_0 + \widetilde{\mathbf{f}}_h(\mathbf{y}_0)h + \tfrac{1}{2}D\widetilde{\mathbf{f}}_h(\mathbf{y}_0)\widetilde{\mathbf{f}}_h(\mathbf{y}_0)h^2 + O(h^3) \\
&= \mathbf{y}_0 + h\mathbf{f}(\mathbf{y}_0) - \tfrac{1}{2}h^2\tfrac{1}{2}D\mathbf{f}(\mathbf{y}_0)\mathbf{f}(\mathbf{y}_0) + \tfrac{1}{2}D\mathbf{f}(\mathbf{y}_0)\mathbf{f}(\mathbf{y}_0)h^2 + O(h^3) \\
&= \mathbf{y}_0 + h\mathbf{f}(\mathbf{y}_0) + O(h^3) = \mathbf{\Psi}^h \mathbf{y}_0 + O(h^3) \,.
\end{aligned}$$

◇

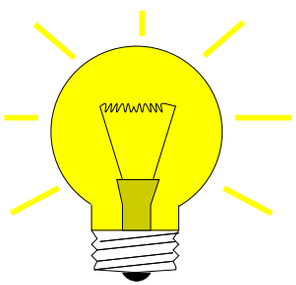
Consistently "implicit assumption": $\mathbf{f}$ "sufficiently smooth" $\Rightarrow$ $\mathbf{\Phi}^t, \mathbf{\Psi}^h$ "sufficiently smooth "

$\mathbf{\Phi}^t \hat{=}$ evolution-operator of the ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$,
$t \mapsto \mathbf{y}(t) \hat{=}$ solution trajectories of $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ to the initial value $\mathbf{y}_0 \in D$.

Goal: formalism of the ad-hoc-construction of a modified equation of order $p+1$ from ex. 4.4.48

Idea: recursive construction of $\widetilde{\mathbf{f}}_h$:

Assumption: discrete evolution $\mathbf{\Psi}^h$ consistent of order p with $\dot{\mathbf{y}} = \mathbf{f}_h(\mathbf{y})$



Ansatz:
$$\boxed{\widetilde{\mathbf{f}}_h = \mathbf{f}_h(\mathbf{y}) + h^p \Delta\mathbf{f}(\mathbf{y})} \qquad (4.4.50)$$

modificatorfunction

Goal:
$$\widetilde{\mathbf{\Phi}}_h^h \mathbf{y} - \mathbf{\Psi}^h \mathbf{y} = O(h^{p+2}) \quad \text{for } h \to 0 \qquad (4.4.51)$$

Explicit assumption, cf. (2.4.13): $\exists$ locally uniformly bounded $\mathbf{d} : D \mapsto \mathbb{R}^d$ with

$$\boldsymbol{\tau}(\mathbf{y}_0,h) := \boldsymbol{\Phi}_h^h\mathbf{y}_0 - \boldsymbol{\Psi}^h\mathbf{y}_0 = \mathbf{d}(\mathbf{y}_0)h^{p+1} + O(h^{p+2}) \quad \text{for} \quad h \to 0\,, \quad \forall \mathbf{y}_0 \in D\,. \tag{4.4.52}$$

Consistency error $\to$ Def. 2.1.11, ($\boldsymbol{\Phi}_h^t \hat{=}$ evolution-operator to $\dot{\mathbf{y}} = \mathbf{f}_h(\mathbf{y})$)

(4.4.51): Determine $\Delta\mathbf{f}$ such that Sol. of $\dot{\widetilde{\mathbf{y}}} = \widetilde{\mathbf{f}}_h(\widetilde{\mathbf{y}})$, $\widetilde{\mathbf{y}}(0) = \mathbf{y}_0$

$$\widetilde{\boldsymbol{\Phi}}_h^h\mathbf{y} - \boldsymbol{\Psi}^h\mathbf{y} = \widetilde{\mathbf{y}}(h) - \mathbf{y}_1 = O(h^{p+2}) \quad \text{for } h \to 0\,. \tag{4.4.53}$$

taylorexpansion of $h = 0$, cf. (2.3.25), used differential equation and chain rule: for $h \to 0$

$$\widetilde{\mathbf{y}}(h) = \mathbf{y}_0 + \sum_{j=1}^{p+1}\frac{h^j}{j!}\widetilde{\mathbf{y}}^{(j)}(0) + O(h^{p+2}) = \mathbf{y}_0 + \sum_{j=1}^{p+1}\frac{h^j}{j!}\frac{d^{j-1}}{dt^{j-1}}\widetilde{\mathbf{f}}_h(\widetilde{\mathbf{y}}(t))_{|t=0} + O(h^{p+2})$$

$$\begin{aligned}
&= \mathbf{y}_0 + h\widetilde{\mathbf{f}}_h(\mathbf{y}_0) + \tfrac{1}{2}h^2 D\widetilde{\mathbf{f}}_h(\mathbf{y}_0)\widetilde{\mathbf{f}}_h(\mathbf{y}_0)\\
&\quad + \tfrac{1}{6}h^3\big(D^2\widetilde{\mathbf{f}}_h(\mathbf{y}_0)(\widetilde{\mathbf{f}}_h(\mathbf{y}_0),\widetilde{\mathbf{f}}_h(\mathbf{y}_0)) + D\widetilde{\mathbf{f}}_h(\mathbf{y}_0)D\widetilde{\mathbf{f}}_h(\mathbf{y}_0)\widetilde{\mathbf{f}}_h(\mathbf{y}_0)\big) + \cdots + O(h^{p+2})\\
&= \mathbf{y}_0 + h\mathbf{f}_h(\mathbf{y}_0) + h^{p+1}\Delta\mathbf{f}(\mathbf{y}_0) + \tfrac{1}{2}h^2 D\mathbf{f}_h(\mathbf{y}_0)\mathbf{f}_h(\mathbf{y}_0)\\
&\quad + \tfrac{1}{6}h^3\big(D^2\mathbf{f}_h(\mathbf{y}_0)(\mathbf{f}_h(\mathbf{y}_0),\mathbf{f}_h(\mathbf{y}_0)) + D\mathbf{f}_h(\mathbf{y}_0)D\mathbf{f}_h(\mathbf{y}_0)\mathbf{f}_h(\mathbf{y}_0)\big) + \cdots + O(h^{p+2})\,,
\end{aligned}$$

since "$O(h^p)$-modified" in (4.4.50), e.g.

$$\begin{aligned}
h^2 D\widetilde{\mathbf{f}}_h(\mathbf{y}_0)\widetilde{\mathbf{f}}_h(\mathbf{y}_0) &= h^2\big(D\mathbf{f}_h(\mathbf{y}_0) + h^p D\Delta\mathbf{f}(\mathbf{y}_0)\big)(\mathbf{f}_h(\mathbf{y}_0) + h^p\Delta\mathbf{f}(\mathbf{y}_0))\\
&= h^2 D\mathbf{f}_h(\mathbf{y}_0)\mathbf{f}_h(\mathbf{y}_0) + O(h^{p+2})\,.
\end{aligned}$$

➢ Observation: taylorexpansion of $t \mapsto \mathbf{\Phi}_h^t \mathbf{y}_0$ around $t = 0$ is contained !

$$\widetilde{\mathbf{y}}(h) = \mathbf{\Phi}_h^h \mathbf{y}_0 + h^{p+1}\Delta\mathbf{f}(\mathbf{y}_0) + O(h^{p+2})$$
$$\overset{(4.4.52)}{=} \mathbf{\Psi}^h \mathbf{y}_0 + h^{p+1}\mathbf{d}(\mathbf{y}_0) + h^{p+1}\Delta\mathbf{f}(\mathbf{y}_0) + O(h^{p+2}) \ .$$

(4.4.53) satisfied by $\boxed{\Delta f(\mathbf{y}) := -\mathbf{d}(\mathbf{y})}$! (4.4.54)

Attempt: series-ansatz for vectorfield of the modified equation:

$$\widetilde{\mathbf{f}}_h(\mathbf{y}) = \mathbf{f}(\mathbf{y}) + h^p\Delta\mathbf{f}_p(\mathbf{y}) + h^{p+1}\Delta\mathbf{f}_{p+1}(\mathbf{y}) + h^{p+2}\Delta\mathbf{f}_{p+2}(\mathbf{y}) + \dots \ . \tag{4.4.55}$$

➢ modified function $\Delta\mathbf{f}_\ell$, $\ell \in \mathbb{N}$, from recursive construction rule

$$(4.4.54) \quad \Rightarrow \quad \Delta\mathbf{f}_\ell(\mathbf{y}) = -\lim_{h\to 0} \frac{\widetilde{\mathbf{\Phi}}_{h,\ell-1}^h \mathbf{y} - \mathbf{\Psi}^h \mathbf{y}}{h^{\ell+1}} \ , \tag{4.4.56}$$

$\widetilde{\mathbf{\Phi}}_{h,\ell}^t \hat{=}$ evolution-operator to the ODE

$$\dot{\widetilde{\mathbf{y}}} = \widetilde{\mathbf{f}}_{h,\ell}(\widetilde{\mathbf{y}}) := \mathbf{f}(\mathbf{y}) + h^p\Delta\mathbf{f}_p(\mathbf{y}) + h^{p+1}\Delta\mathbf{f}_{p+1}(\mathbf{y}) + h^{p+2}\Delta\mathbf{f}_{p+2}(\mathbf{y}) + \dots + h^\ell\Delta\mathbf{f}_\ell(\mathbf{y}) \ . \tag{4.4.57}$$

Explanation for (4.4.56): formula for leading term in the taylorexpansion of the consistency-error:

$$\boldsymbol{\tau}(\mathbf{y},h) = \boldsymbol{\Phi}^h\mathbf{y} - \boldsymbol{\Psi}^h\mathbf{y} = \mathbf{d}(\mathbf{y})h^{p+1} + O(h^{p+2}) \quad \Rightarrow \quad \mathbf{d}(\mathbf{y}) = \lim_{h\to 0} \frac{\boldsymbol{\Phi}^h\mathbf{y} - \boldsymbol{\Psi}^h\mathbf{y}}{h^{p+1}} \,.$$

Remark 4.4.58 (Computation of the modificatorfunctions $\Delta\mathbf{f}_j$ by computeralgebra).

MAPLE-Code: Computation of Δf_j

```
fcn := y ->f(y) :
N :=q :
fcoe [1] := fcn(y):
for n from 2 by 1 to N do
  modeq := sum(h^j*fcoe [j+1], j=0..n-2):
  diffy [0] := y:
  for i from 1 by 1 to n do
    diffy [i] := diff(diffy[i-1],y)*modeq:
  od:
  ytilde := sum(h^k*diffy[k]/k!, k=0..n):
  res := ytilde-y-h*fcn(y):
  tay := convert(series(res,h=0,n+1),polynom):
  fcoe [n] := -coeff(tay,h,n):
```

OSM:

Explicite euler-method

◁ MAPLE-code [18]:

Computation of modificatorfunctions Δf_ℓ for scalar ODE $\dot{y} = f(y)$.

Output of the series (4.4.55)

Example 4.4.59 (modificator for simple OSM).

Scalar differential equation: $\dot{y} = y^2$ $\rightarrow$ Ex. 1.3.11

- explicit euler-method (1.4.2): $y_1 = y_0 + hf(y_0)$

$$\widetilde{f}(y) = y^2 \underbrace{- h\, y^3}_{-\Delta f_1} \underbrace{+h^2\, 3/2\, y^4}_{\Delta f_2} \underbrace{-h^3\, 8/3\, y^5}_{-\Delta f_3} \underbrace{+h^4\, \frac{31}{6}\, y^6}_{\Delta f_4} \underbrace{-h^5\, \frac{157}{15}\, y^7}_{-\Delta f_5}$$
$$\underbrace{+ h^6\, \frac{649}{30}\, y^8}_{\Delta f_6} \underbrace{-h^7\, \frac{9427}{210}\, y^9}_{-\Delta f_7} \underbrace{+h^8\, \frac{19423}{210}\, y^{10}}_{\Delta f_8} \underbrace{-h^9\, \frac{6576}{35}\, h^9 y^{11}}_{-\Delta f_9} + O(h^{10})\ .$$

- implicit euler-method (1.4.13): $y_1 = y_0 + hf(y_1)$

(In MAPLE code: `res := ytilde-y-h*fcn(ytilde)`)

$$\widetilde{f}(y) = y^2 + hy^3 + 3/2\, h^2 y^4 + 8/3\, h^3 y^5 + \frac{31}{6}\, h^4 y^6 + \frac{157}{15}\, h^5 y^7 + \frac{649}{30}\, h^6 y^8$$
$$+ \frac{9427}{210}\, h^7 y^9 + \frac{19423}{210}\, h^8 y^{10} + \frac{6576}{35}\, h^9 y^{11} + O(h^{10})$$

- implicit midpoint-rule (1.4.19): $y_1 = y_0 + hf(\frac{1}{2}(y_0 + y_1))$

(In MAPLE code: `res := ytilde-y-h*fcn(0.5*(y+ytilde))`)

$$\widetilde{f}(y) = y^2 + \frac{1}{4}h^2y^4 + \frac{1}{8}h^4y^6 + 0.057291667\,h^6y^8 + 0.02343750000\,h^8y^{10} + O(h^{10}) \;.$$

☞ only *even* powers of h, cf. proof of Thm. 2.1.29, Thm. 2.4.22

◇

power series (in h) $\sum_{k=1}^{\infty} h^k \Delta f_k(\mathbf{y})$ possibly divergent

$\forall h > 0$ ($\leftrightarrow$ convergence-radius $= 0$)

interpretation of (4.4.55) as asymptotic expansion of $\widetilde{\mathbf{f}}_h$, see def. 2.4.7

Example 4.4.60 (Meaning of the modified equation of low order).

- Initial value problem for logistic differential equation, see Ex. 1.2.1

$$\dot{y} = \lambda y(1-y) \quad , \quad y(0) = 0.01 \;.$$

- OSM: explicite euler-method (1.4.2), cf. Ex. 1.4.9, modificatorfunctions from (4.4.55)

$$\Delta f_1(y) = \lambda^2\big(-1/2\,y + 3/2\,y^2 - y^3\big)\ , \quad \Delta f_2(y) = \lambda^3\big(-\frac{11}{6}\,y^2 + 3\,y^3 - 3/2\,y^4 + 1/3\,y\big)\ .$$

- OSM: implicite euler-method (1.4.13), cf. ex. 1.4.15, modificatorfunctions from (4.4.55)

$$\Delta f_1(y) = \lambda^2\big(1/2\,y - 3/2\,y^2 + y^3\big)\ , \quad \Delta f_2(y) = \lambda^3\big(-\frac{11}{6}\,y^2 + 3\,y^3 - 3/2\,y^4 + 1/3\,y\big)$$

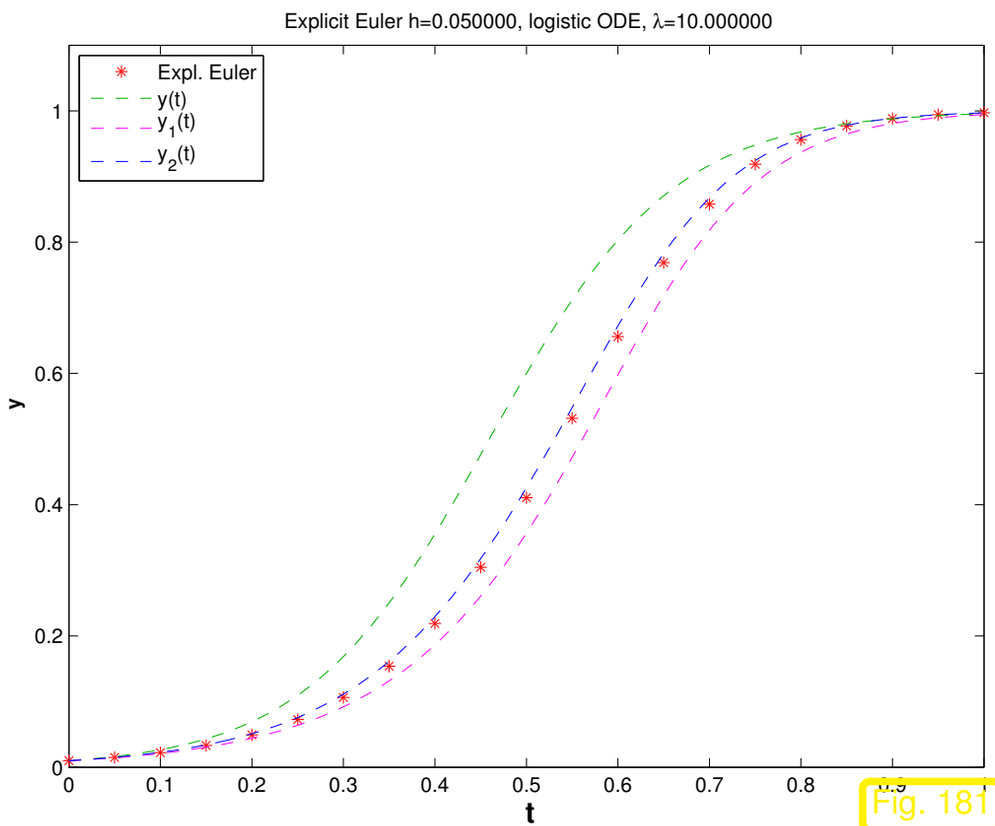


Fig. 181

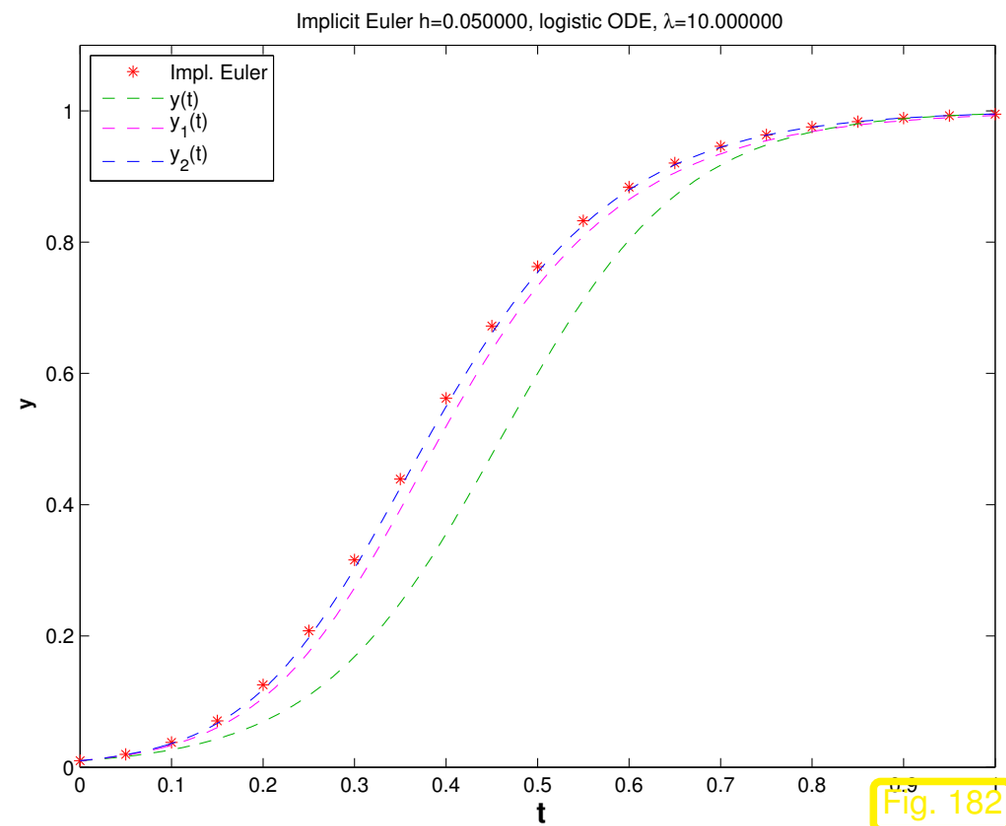


Fig. 182

The euler-method for $y = f(y)$ yield a better approximation for the solution of

$$\dot{y} = \widetilde{f}_{1,h}(y) = f(y) + h\Delta f_1(y) \quad \text{und} \quad \dot{y} = \widetilde{f}_{1,h}(y) = f(y) + h\Delta f_1(y) + h^2\Delta f_2(y) \ .$$

► Consider cropped modified equation !

◇

Lemma 4.4.61 ("cropped" modified equation)**.**
Mit modificatorfunctions $\Delta\mathbf{f}_i$ according (4.4.56) *for the ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ and the OSM with discrete evolution $\mathbf{\Psi}^h$ (of the consistency-order p) as defined above, is*

$$\dot{\widetilde{\mathbf{y}}} = \widetilde{\mathbf{f}}_{h,\ell}(\widetilde{\mathbf{y}}) := \mathbf{f}(\widetilde{\mathbf{y}}) + h^p\Delta\mathbf{f}_p(\widetilde{\mathbf{y}}) + h^{p+1}\Delta\mathbf{f}_{p+1}(\widetilde{\mathbf{y}}) + \cdots + h^\ell\Delta\mathbf{f}_\ell(\widetilde{\mathbf{y}}) \ ,$$

a modified equation of order $\ell + 1$, $\ell > p$ ($\rightarrow$ def. 4.4.47)

Proof. The proof ensues by recursive construction of $\Delta\mathbf{f}_\ell$, see (4.4.52), (4.4.53), (4.4.54).

4.4.4 Modified equation: error analysis

In the sense of backward error analysis ($\rightarrow$ Rem. 4.4.43) of the *longtime behavior* of the one-step-method we have to study:

- Is there any (structure-preserving) modified equation, which solution stays close to the numerical solution (gridfunction $(\mathbf{y}_k)_k$) for long time?
- If yes, is the right hand side of this modified equation close to the right hand side of the initial equation?

Strategy: What do we want **?**

Lemma 4.4.61: *Family* modified equations $\dot{\widetilde{\mathbf{y}}} = \widetilde{\mathbf{f}}_{h,\ell}(\widetilde{\mathbf{y}})$, "consistent with the OSM" $\mathbf{y}_{k+1} = \boldsymbol{\Psi}^h \mathbf{y}_k$, that means

$$\text{Consistency error} \qquad \boldsymbol{\tau}(\mathbf{y},h) := \boldsymbol{\Psi}^h \mathbf{y} - \widetilde{\boldsymbol{\Phi}}^h_{h,\ell}\mathbf{y} \to 0 \quad \text{for } h \to 0 \,.$$

Reminder to the proof of *Convergencetheorem for OSM*, Thm. 2.1.19

(cf. proof of Thm. 2.1.26 and the "discrete Gronwall-lemma" lemma 2.1.20)

$$\|\mathbf{y}_k - \widetilde{\mathbf{y}}(kh)\| \leq \frac{1}{h} \max_{j=0,\dots,k-1} \left\|\boldsymbol{\tau}(\mathbf{y}_j, h)\right\| \frac{\exp(Lhk) - 1}{L} \,. \tag{4.4.62}$$

($L > 0$: Lipschitz-constant of the increment-function of the OSM, $\widetilde{\mathbf{y}} \hat{=}$ Solution of the modified equation)

Exponential growth of the constant in (4.4.62) for $hk \to \infty$!

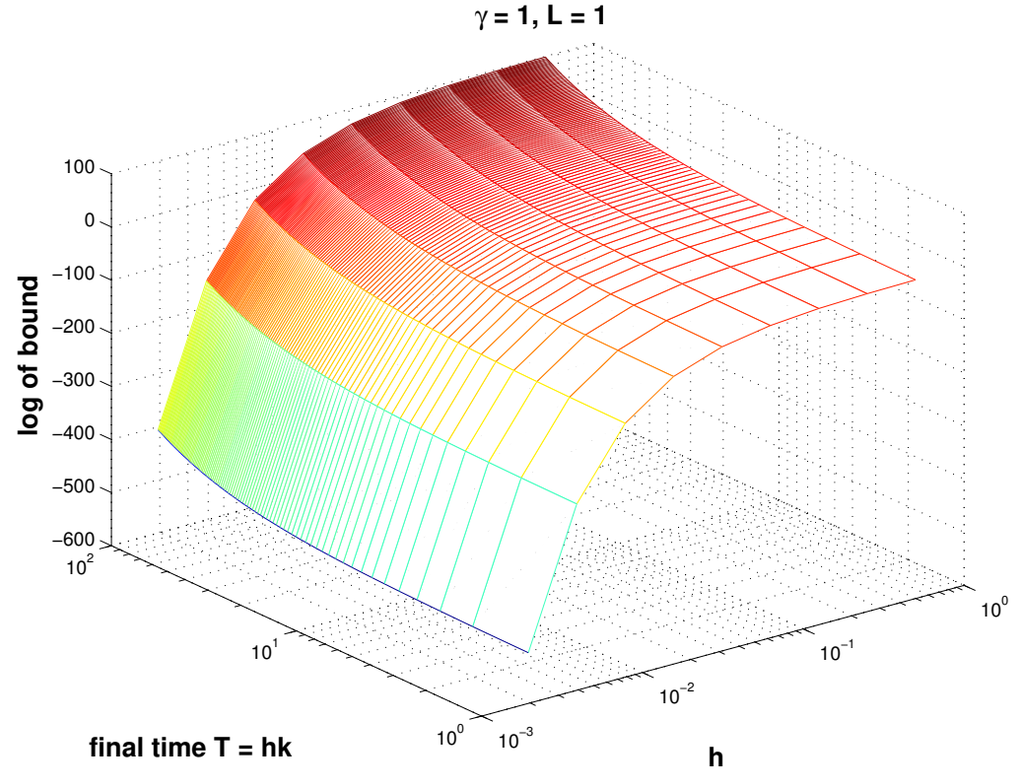
($\left\|\boldsymbol{\tau}(\mathbf{y}_j, h)\right\| = O(h^{\ell+2})$ does not yield any useful estimate for *longtime-integration*)

But:

If consistency error "exponentially small"

$$\boxed{\|\boldsymbol{\tau}\| \leq Ch\exp(-\gamma/h)\ , \quad \gamma > 0} \ . \qquad (4.4.63)$$

$$\|\mathbf{y}_k - \widetilde{\mathbf{y}}(kh)\| \leq C\exp(-\gamma/h + hkL)\ . \qquad (4.4.64)$$



Behavior of the bound from (4.4.64)

P. Grohs

rev 63606, February 17, 2014

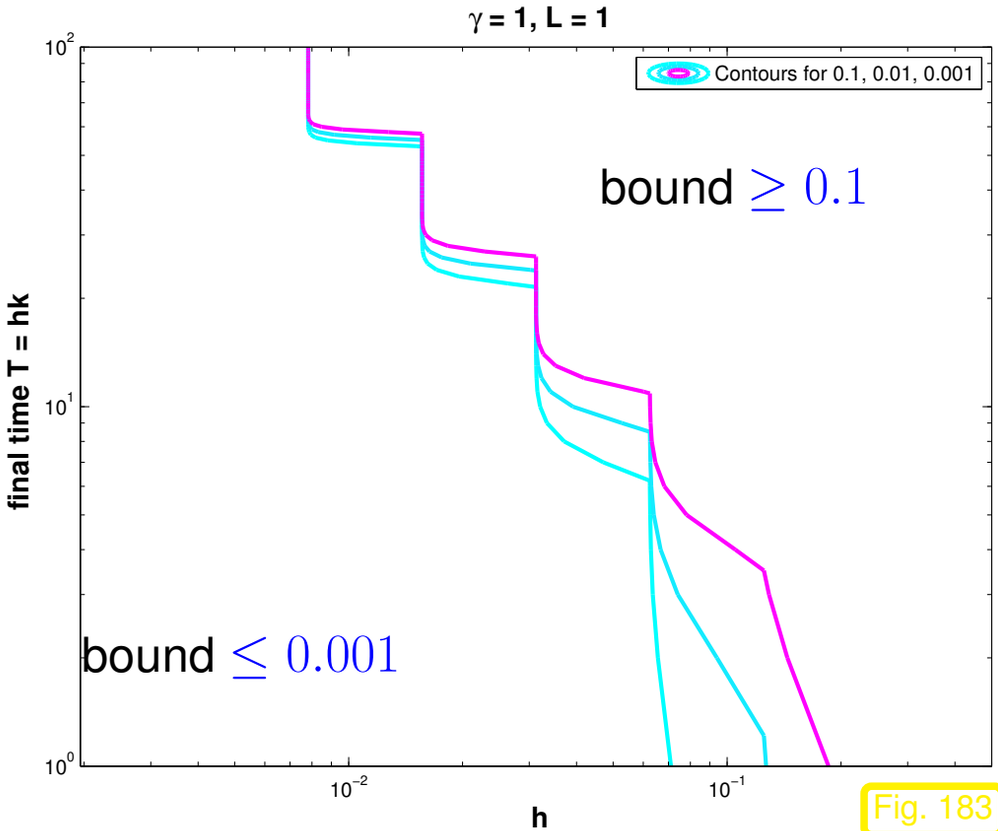


Fig. 183

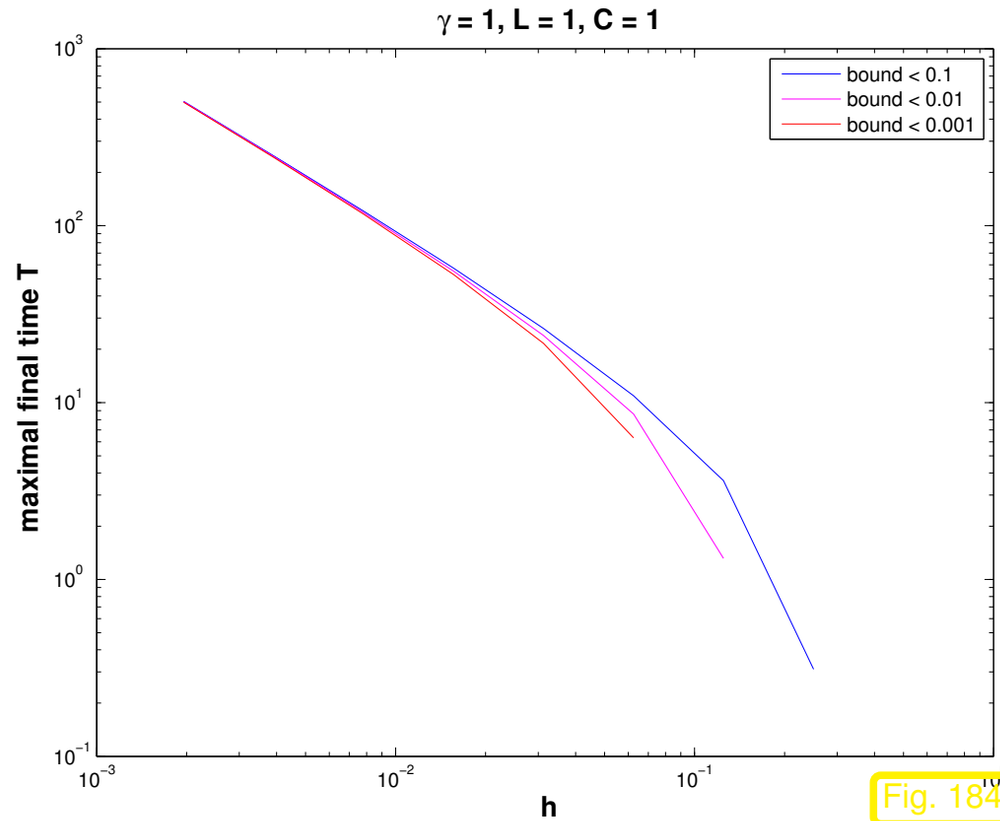


Fig. 184

P. Grohs

rev 63606, February 17, 2014

From (4.4.64) we see that:

$$h < \frac{\gamma}{TL - \log(\tau/C)} \quad \Rightarrow \quad \|\mathbf{y}_k - \widetilde{\mathbf{y}}(kh)\| \leq \tau \quad \text{for} \quad 0 \leq kh \leq T\,.$$

numerical solution $\mathbf{y}_k$ stays a long time near the trajectories $t \mapsto \widetilde{\mathbf{y}}(t)$

(precise discussion in Rem. 4.4.85)

We are free in the choice of the cropping index ℓ **!**

Question: What is the best cropped modified equation**?**

To answer we need a concept from function theory!

Analyticity condition

for each compact $K \subset D$ there is a $R = R(K) > 0$, such that $\mathbf{f}(\mathbf{y})$ in each $\mathbf{y} \in K$ in each component of $\mathbf{y}$ has a power series expansion with convergence radius $> R$.

$\Leftrightarrow$ $\mathbf{f}$ is holomorphic in D

P. Grohs

rev 63606, February 17, 2014

Explanation: power series expansion around $\mathbf{y} = (y_1, \dots, y_d)^T$ in the j. component

$$\mathbf{f}(y_1, \dots, y_{j-1}, y, y_{j+1}, \dots, y_d) = \sum_{k=0}^{\infty} \mathbf{a}_k(\mathbf{y})(y - y_j)^k \quad \text{for} \quad |y - y_j| < R \,.$$

Example 4.4.65 (Analyticity condition for hamiltonian dfferential equation).

$$\mathbf{f}(\mathbf{y}) = \mathbf{J}^{-1}\,\mathbf{grad}\,H(\mathbf{y}) \text{ holomorphic in } D \quad \Leftrightarrow \quad H(\mathbf{y}) \text{ holomorphic in } D$$

(with the same lower bound R for convergence radius on compact regions)

- mathematical pendulum, Ex. 4.4.3: $H(p,q) = \frac{1}{2}p^2 - \cos q$
  ➢ $D = \mathbb{R}^2$, $R = \infty$ (whole function !)
- spring pendulum, Ex. 4.4.35: $H(\mathbf{p},\mathbf{q}) = \frac{1}{2}\left\|\mathbf{p}\right\|^2 + \frac{1}{2}(\left\|\mathbf{q}\right\| - 1)^2$
  ➢ $D = \mathbb{R}^4$, $R = \mathrm{dist}(K, \{\mathbf{q} = 0\})$ (H not holomorphic in $\mathbf{q} = 0$)

P. Grohs

rev 63606, February 17, 2014

Observe: $H(\mathbf{p},\mathbf{q})$ analytical in a neighborhood of physical useful trajectories!

◇

✎ Notation: $\widetilde{\mathbf{\Phi}}^t_{h,\ell} \hat{=}$ evolution operator for $\dot{\widetilde{\mathbf{y}}} = \widetilde{\mathbf{f}}_{h,\ell}(\widetilde{\mathbf{y}})$, cf. lemma 4.4.61

Theorem 4.4.66 (Consistency error estimation for cropped modified equations)**.**
Let $\mathbf{\Psi}^h$ be the discrete evolution of a $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ consistent (partitioned) runge-kutta-one-step method. Because of the analytizity condition there are for each compact $K \subset D$ constants $C_1, C_2 > 0$ and a $h_0 \in]0, \infty]$ such that

$$\left\| \mathbf{\Psi}^h \mathbf{y} - \widetilde{\mathbf{\Phi}}^h_{h,\ell} \mathbf{y} \right\| \leq C_1 h (C_2(\ell+1)h)^{\ell+1} \quad \forall \mathbf{y} \in K, \quad \forall \ell \in \mathbb{N}\,, \quad \forall |h| \leq h_0\,. \tag{4.4.67}$$

Tool for the proof: differential equation in $\mathbb{C}$ $\rightarrow$ thm. 2.2.82

P. Grohs

rev 63606, February 17, 2014

Lemma 4.4.68. *If f is holomorphic in a neighborhood of $B_\rho(0)$, $|f(z)| \leq M$ for all $z \in B_\rho(0)$ and $f(0) = \cdots = f^{(p)}(0) = 0$, $p \in \mathbb{N}_0$, then we have*

$$|f(z)| \leq M \left|\frac{z}{\rho}\right|^{p+1} \quad \forall z \in B_\rho(0)\,.$$

Proof. On $B_\rho(0)$

$$f(z) = z^{p+1} \underbrace{\sum_{j=0}^{\infty} a_j z^j}_{=:g(z)} \quad , \quad |g(z)| \leq \frac{M}{\rho^{p+1}} \text{ for } |z| = \rho \ .$$

g holomorphic on $B_\rho(0)$ $\Rightarrow |g|$ attains maximum on the boundary $|z| = \rho$ (maximum principle). □

Boundedness on a zeroset of a holomorphic function f with $|f(z)| = O(|z|^{p+1})$ is enough, to characterize the decreasing behavior for $z \to 0$!

☞ for scalar case $d = 1$, $\dot{y} = f(y)$, $D =]a, b[\subset \mathbb{R}$ interval,

☞ for explicit euler-method (1.4.2): $\Psi^h y = y + hf(y)$ (Proof from S. Reich 1999, see [35, Thm. 2])

Assumption:

f holomorphic in neighborhood of

$$B_R(D) := \{z \in \mathbb{C} \colon \exists x \in D : \ |z - x| \leq R\}$$

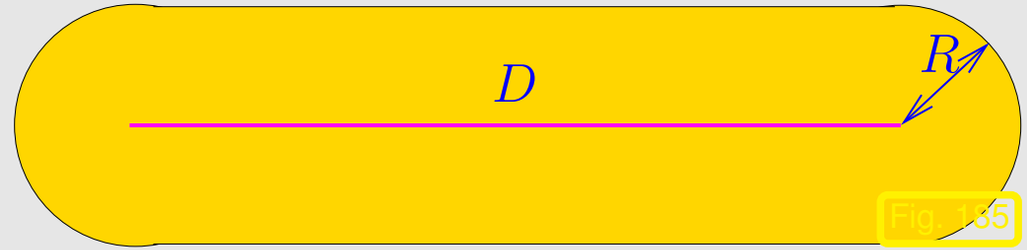


Fig. 185

Estimation of the consistency error $\widetilde{\Phi}^h_{h,\ell}(y) - \Psi^h(y)$ for modified equation of order $\ell+1$ ($\rightarrow$ def. 4.4.47)
$\Updownarrow \leftarrow$ (4.4.54)
Estimation of the modificator functions Δf_ℓ !

P. Grohs

rev 63606, February 17, 2014

Step I: Estimation for modificator function Δf_1 (also to demonstrate the technique)

interpretation of $\dot{y} = f(y)$ as differential equation in $\mathbb{C}$ $\rightarrow$ thm. 2.2.82 : $\Rightarrow$ solution $t \mapsto y(t)$ analytical (in neighborhood of 0) $\Rightarrow$ can be continued by $\mathbb{C}$
$\Rightarrow$ evolution $\Phi^t : B_R(D) \mapsto \mathbb{C}$ holomorphic (for sufficiently small $|t|$)

▶ In the following: consider complex "timestepwidth" $h \in \mathbb{C}$, $0 < \alpha < 1$ chosen fixed.

From the estimation for path integrals in the complex plane

$$M := \max_{z\in B_R(D)} |f(z)| \quad \Rightarrow \quad |\Phi^h z - z| = \left|\int_0^h f(\Phi^\tau z)\,\mathrm{d}\tau\right| \le M|h| \;, \quad \begin{array}{c}\forall z \in B_{\alpha R}(D)\\ |h| \text{ small.}\end{array} \tag{4.4.69}$$

$$\Rightarrow \quad |\Psi^h z - z| = |hf(z)| \le M|h| \;.$$

Upper bound for $|h|$:

If $z \in B_{\alpha R}(D)$ & $|h| \le (1-\alpha)\frac{R}{M}$, $0 \le \alpha < 1$ the trajectory stays $\xi \mapsto \Phi^{\xi h} z$, $0 \le \xi \le 1$, in $B_R(D)$!

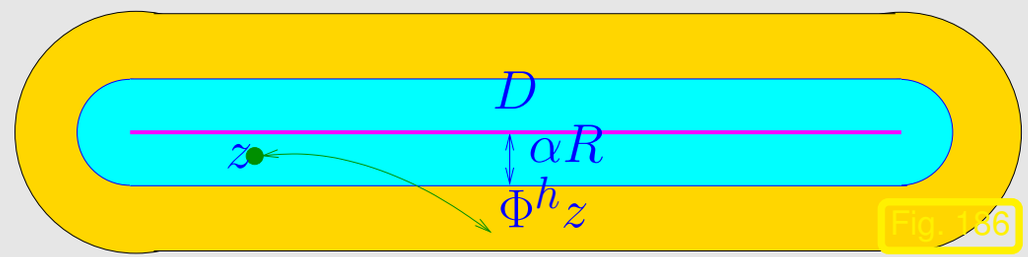


Fig. 136

$$|h| \le h_1 := \frac{(1-\alpha)R}{M} \quad \Rightarrow \quad \begin{array}{l}|\Phi^h y - y| \le (1-\alpha)R\;,\\ |\Psi^h y - y| \le (1-\alpha)R\;.\end{array} \quad \forall y \in B_{\alpha R}(D)$$

$$\Rightarrow \quad \left(|h| \le \frac{(1-\alpha)R}{M} \quad \Rightarrow \quad |\Phi^h y - \Psi^h y| \le 2(1-\alpha)R \quad \forall y \in B_{\alpha R}(D)\right) \;. \tag{4.4.70}$$

P. Grohs

rev 63606, February 17, 2014

Application of lemma 4.4.68 on $g(h) = \Phi^h y - \Psi^h y$:

- g holomorphic in $B_{(1-\alpha)R/M}(0)$
- g bounded by $2(1-\alpha)R$ on $B_{(1-\alpha)R/M}(0)$

$$\begin{aligned} \Rightarrow \quad |\Phi^h y - \Psi^h y| &\leq 2(1-\alpha)R|h|^2 \left(\frac{(1-\alpha)R}{M}\right)^{-2} \\ &\leq 2M|h|^2 \left(\frac{M}{(1-\alpha)R}\right) \quad \forall y \in B_{\alpha R}(D) \ , \end{aligned} \tag{4.4.71}$$

da $g(h) = O(h^2)$ (euler-method consistency order 1), such that $g(0) = g'(0) = 0$.

$$\overset{(4.4.56)}{\Rightarrow} \quad |\Delta f_1(y)| = \left| \lim_{h\to 0} \frac{\Phi^h y - \Psi^h y}{h^2} \right| \overset{(4.4.71)}{\leq} 2M \left(\frac{M}{(1-\alpha)R}\right) \quad \forall y \in B_{\alpha R}(D) \ . \tag{4.4.72}$$

We have seen how we can get with the analyticity assumption on the right hand side f an estimation for the first modificator function. Required is a bound for f in a compact neighborhood $B_R(D) \subset \mathbb{C}$.

The recursive construction of the modificator function according to (4.4.56) suggests the following approach:

① Using estimations for the modificator functions Δf_j, $1 \leq j \leq \ell$, derive an estimation for the right hand side $\widetilde{f}_{h,\ell}$ the modified equation (4.4.57) from lemma 4.4.61. As all modificator functions also $\widetilde{f}_{h,\ell}$ will be analytical in a neighborhood of D.

② Use the bound for $\widetilde{f}_{h,\ell}$, to estimate with the same techniques as above for Δf_1 the next modificator function $\Delta f_{\ell+1}$.

③ Continue with ①

▶ Recursive estimation $\longleftrightarrow$ proof by induction

Challenge: Formulation of a suitable induction hypothesis cf. (4.4.72).

Step II. Proof by induction: induction assumption: There is ℓ-*independent* $b > 0$, $c > 0$, such that

$$\forall l \in \mathbb{N}: \quad \max_{y \in B_{\alpha R}(D)} |\Delta f_\ell(y)| \leq bM \left(\frac{c\ell M}{(1-\alpha)R} \right)^\ell \quad \forall y \in B_{\alpha R}(D) \,, \quad \forall 0 \leq \alpha < 1 \,. \tag{4.4.73}$$

the constants b, c will be fixed later.

inductions basis "$\ell = 0$" $\Leftrightarrow$ (4.4.72)

induction step "$\ell \Rightarrow \ell+1$": ($0<\alpha<1$ fixed!)

$$
\begin{aligned}
|\widetilde{f}_{h,\ell}(y)| \leq & |f(y)| + |h||\Delta f_1(y)| + |h|^2|\Delta f_2(y)| + \cdots + |h|^\ell|\Delta f_\ell(y)| \\
\leq & M + |h|\frac{2M}{(1-\alpha)R} + bM\sum_{j=2}^{\ell}|h|^j\left(\frac{jcM}{(1-\alpha)R}\right)^j , \quad \begin{array}{l}\forall y \in B_{\alpha R}(D)\ , \\ \forall 0<\alpha<1\ .\end{array}
\end{aligned}
$$

aus (4.4.72) after induction assumption (4.4.73)

Required: Bound for $|\widetilde{f}_{h,\ell}(y)|$ in a *neighborhood* of $B_{\alpha R}(D)$

Idea: "$\forall\alpha$" in (4.4.73) ➢ use freedom in the choice of α !

$$
\alpha^* := \alpha + \delta(1-\alpha) \in]\delta, 1[\quad \Rightarrow \quad 1-\alpha^* = (1-\alpha)(1-\delta) \quad , \quad \alpha^* > \alpha\ .
$$

$$
\Rightarrow \quad \max_{y\in B_{\alpha^* R}(D)} |\widetilde{f}_{h,\ell}(y)| \leq M + |h|\frac{2M}{(1-\alpha)(1-\delta)R} + bM\sum_{j=2}^{\ell}\left(\frac{jcM|h|}{(1-\alpha)(1-\delta)R}\right)^j .
$$

experiment: Simplify by constraint $|h|$: $\boxed{|h| \leq h_\ell := \frac{(1-\alpha)R}{(\ell+1)cM}}$

$$
\Rightarrow \quad \max_{y\in B_{\alpha^* R}(D)} |\widetilde{f}_{h,\ell}(y)| \leq M\left(1 + \frac{2}{c(\ell+1)(1-\delta)} + b\sum_{j=2}^{\ell}\left(\frac{j}{(\ell+1)(1-\delta)}\right)^j\right) . \qquad \textbf{(4.4.74)}
$$

P. Grohs

rev 63606, February 17, 2014

Reminder to our goal (4.4.73) for "$\ell \leftarrow \ell+1$". Because of

$$\Delta \mathbf{f}_{\ell+1}(y) = -\lim_{h\to 0} \frac{\widetilde{\Phi}^h_{h,\ell} y - \Psi^h y}{h^{\ell+2}} \,, \tag{4.4.56}$$

we have to show:

$$|\widetilde{\Phi}^h_{h,\ell} y - \Psi^h y| \le |h|^{\ell+2} bM \underbrace{\Big(\frac{c(\ell+1)M}{(1-\alpha)R}\Big)}_{=h_\ell^{-1}\,!}{}^{\ell+1} |h|^{\ell+2} bM h_\ell h_\ell^{-(\ell+2)} \quad \forall y \in B_{\alpha R}(D) \tag{4.4.75}$$

Note: $\boxed{\text{(4.4.75)} \quad \Rightarrow \quad \text{claim of theorem with } C_1 = bM,\ C_2 = \frac{cM}{(1-\alpha)R}\ \textbf{!}}$

Note: *per construction*, lemma 4.4.61: $\boxed{\widetilde{\Phi}^h_{h,\ell} y - \Psi^h y = O(h^{\ell+2}) \text{ for } h \to 0}$

Lemma 4.4.68 $\Rightarrow$ Since $h \mapsto \widetilde{\Phi}^h_{h,\ell} y - \Psi^h y$ analytical, it is enough to prove

$$|\widetilde{\Phi}^h_{h,\ell} y - \Psi^h y| \le h_\ell bM \quad \forall h \in B_{h_\ell}(0)\,, \quad \forall y \in B_{\alpha R}(D)\,. \tag{4.4.76}$$

Then triangle inequality as in (4.4.70) & Estimation analog to (4.4.69):

$$|\widetilde{\Phi}^h_{h,\ell} y - y| \le |h| \max_{y \in B_{\alpha^* R}(D)} |\widetilde{f}_{h,\ell}(y)| \quad \forall y \in B_{\alpha R}(D),\ |h| \text{ "sufficiently small".} \tag{4.4.77}$$

What do we need ? ($|\Psi^h y - y| \leq M|h|$ as above)

- $\max\limits_{y \in B_{\alpha^* R}(D)} |\widetilde{f}_{h,\ell}(y)| \leq (b-1)M$,
- $h_\ell \cdot \max\limits_{y \in B_{\alpha^* R}} |\widetilde{f}_{h,\ell}(y)| \leq \delta(1-\alpha)R$, such that the trajectory $z \mapsto \widetilde{\Phi}^z_{h,\ell} y$ in $B_{\alpha^* R}(D)$ stays, if $|z| \leq h_\ell$ (Note: $B_{\alpha R}(D) \subset B_{\alpha^* R}(D)$).

To this end we have to choose the parameters in (4.4.74) appropriate!

We are "free" in the choice of $\delta \in]0,1[$! $\quad\rightarrow\quad \boxed{\delta := \frac{b-1}{c} \cdot \frac{1}{\ell+1}} \quad \Rightarrow \quad h_l(b-1)M = \delta(1-\alpha)R$

$$\overset{(4.4.74)}{\Rightarrow} \quad \max_{y \in B_{\alpha^* R}(D)} |\widetilde{f}_{h,\ell}(y)| \leq M \underbrace{\left(1 + \frac{2}{c(\ell+1)-b+1} + b\sum_{j=2}^{\ell}\left(\frac{jc}{c(\ell+1)-b+1}\right)^j\right)}_{=:\Gamma(b,c,\ell)} .$$

Question: Are there $b, c > 0$ ($b - 1 < 2c$) such that $\boxed{\max\limits_{\ell \in \mathbb{N}} \Gamma(b,c,\ell) \leq b-1}$?

For the proof of thm. 4.4.66: behavior of

$$\Gamma(b,c,\ell) := 1 + \frac{2}{c(\ell+1)-b+1} + b\sum_{j=2}^{\ell}\left(\frac{jc}{c(\ell+1)-b+1}\right)^j :$$

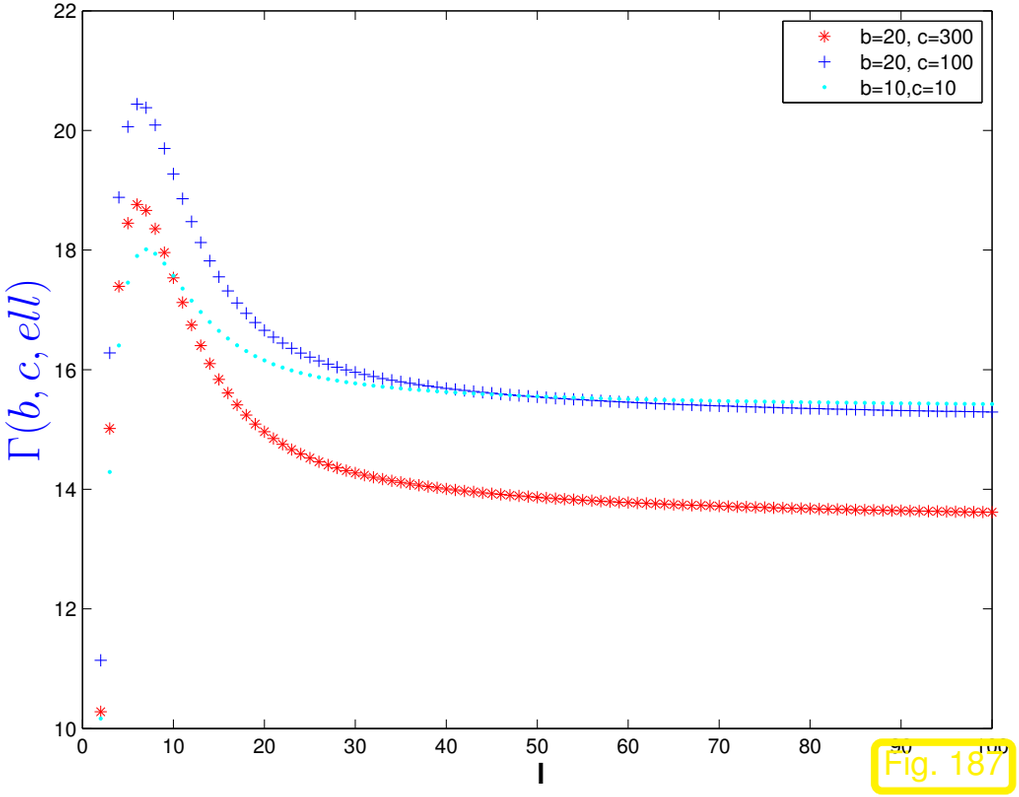


Fig. 187

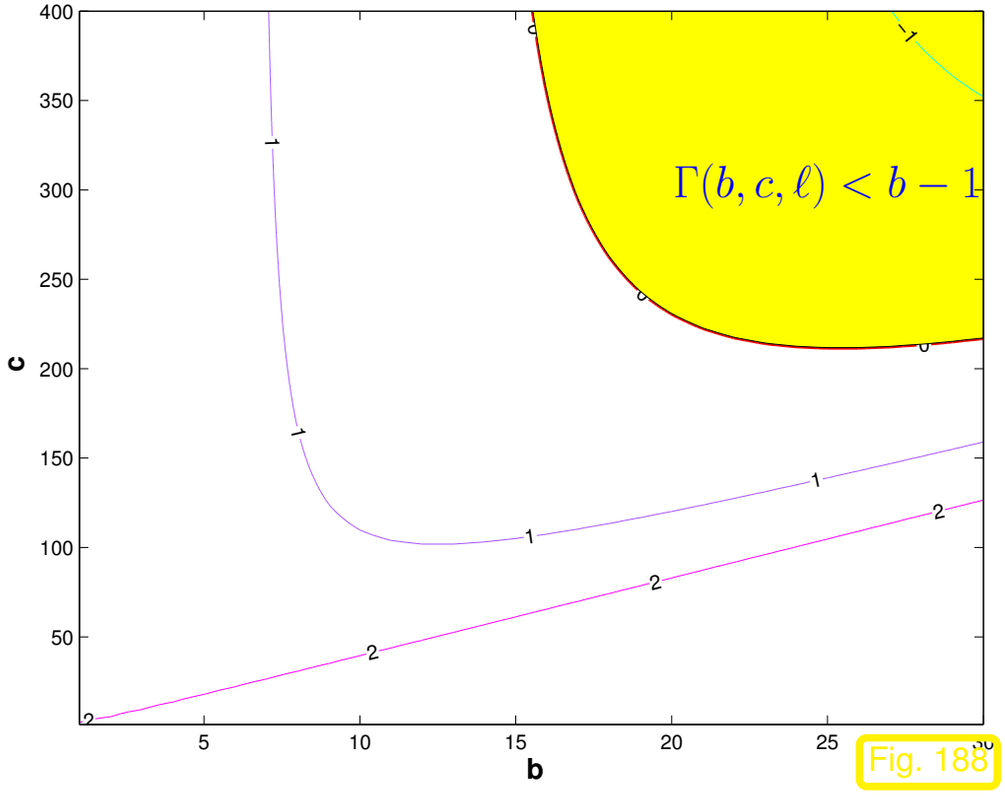


Fig. 188

left plot: $\ell \mapsto \Gamma(b,c,\ell)$ ➢ unique maximum for small ℓ

right plot: contures of $(b,c) \mapsto \max_\ell \Gamma(b,c,\ell) - b + 1$

▶ From the plots we see: possible choice $c = 300$, $b = 20$ $\forall \ell$.

Then continue as outlined before, see (4.4.76), (4.4.77):

$$\begin{aligned} &\Rightarrow \quad |\widetilde{\Phi}^h_{h,\ell} y - y| \le M(b-1)|h| \,, \\ &\Rightarrow \quad |\widetilde{\Phi}^h_{h,\ell} y - \Psi^h y| \le bM|h| \end{aligned} \quad \text{for } |h| \le h_\ell \,, \quad \forall y \in B_{\alpha R}(D) \,, \tag{4.4.78}$$

Note: $\widetilde{\Phi}^h_{h,\ell} y - \Psi^h y = O(h^{\ell+2})$ by construction of the modificator functions and holomorphic in a neighborhood of 0. With formula for h_ℓ, o.B.d.A. $0 < h_\ell < 1$,

$$\overset{\text{Lemma 4.4.68}}{\Rightarrow} \quad |\widetilde{\Phi}^h_{h,\ell} y - \Psi^h y| \le bM \left(\frac{|h|}{h_\ell}\right)^{\ell+2} \le bM|h|^{\ell+2} \left(\frac{c(\ell+1)M}{(1-\alpha)R}\right)^{\ell+1} . \tag{4.4.79}$$

With (4.4.56) the induction hypothesis for $\ell + 1$.

Claim of the theorem with $C_1 = bM$, $C_2 = \frac{cM}{R}$ (Case $\alpha = 0$) follows also by (4.4.79) □

MATLAB: /scratch/users/hiptmair/NOSAVE/numpde/Numcourses/NumerischeMathematik/MatlabRevised/modeqmean

Behavior of the bound from Thm. 4.4.66 ▷

possible divergence of the asymptotic development (4.4.55) manifests itself in $C_1 h(C_2(\ell+1)h)^{\ell+1} \to \infty$ for $\ell \to \infty$.

optimal break off index:

$$\ell_{\text{opt}} \approx \left[\frac{1}{C_2 e h}\right] . \tag{4.4.80}$$

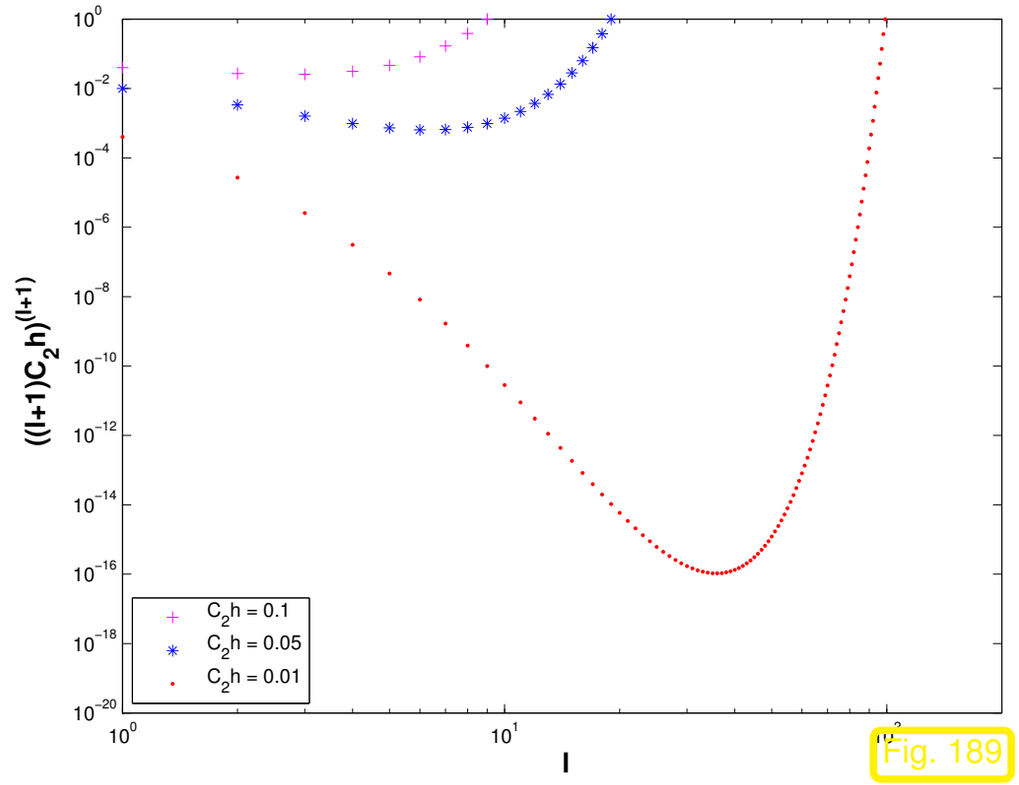


Fig. 189

P. Grohs

rev 63606, February 17, 2014

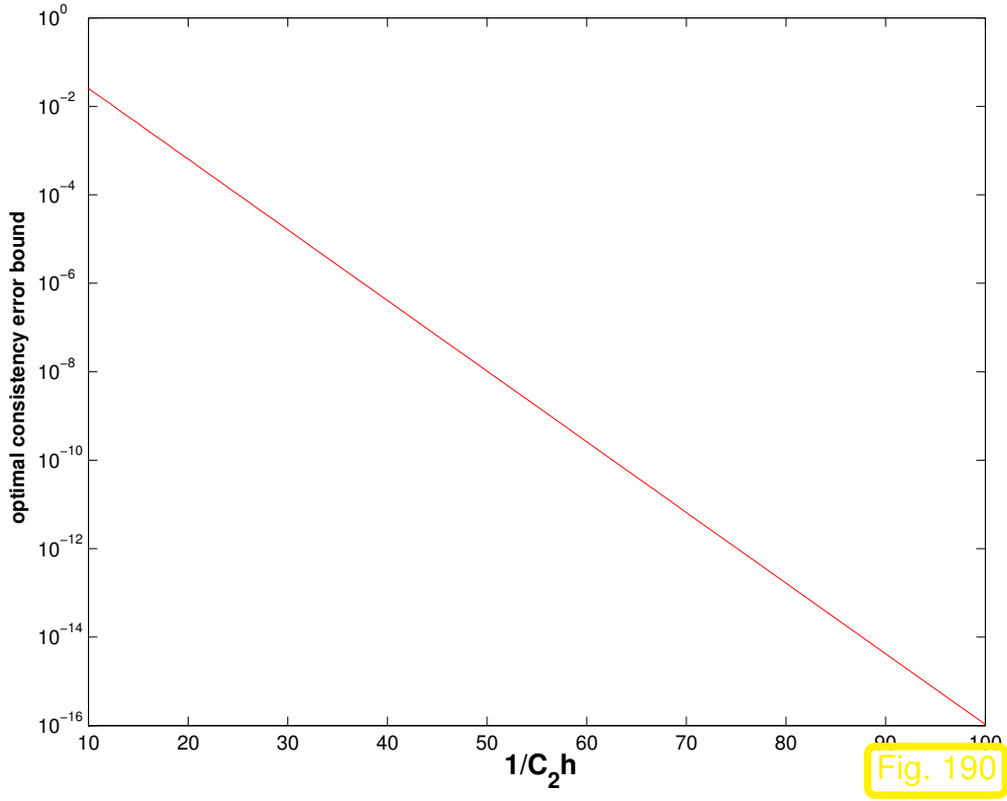


Fig. 190

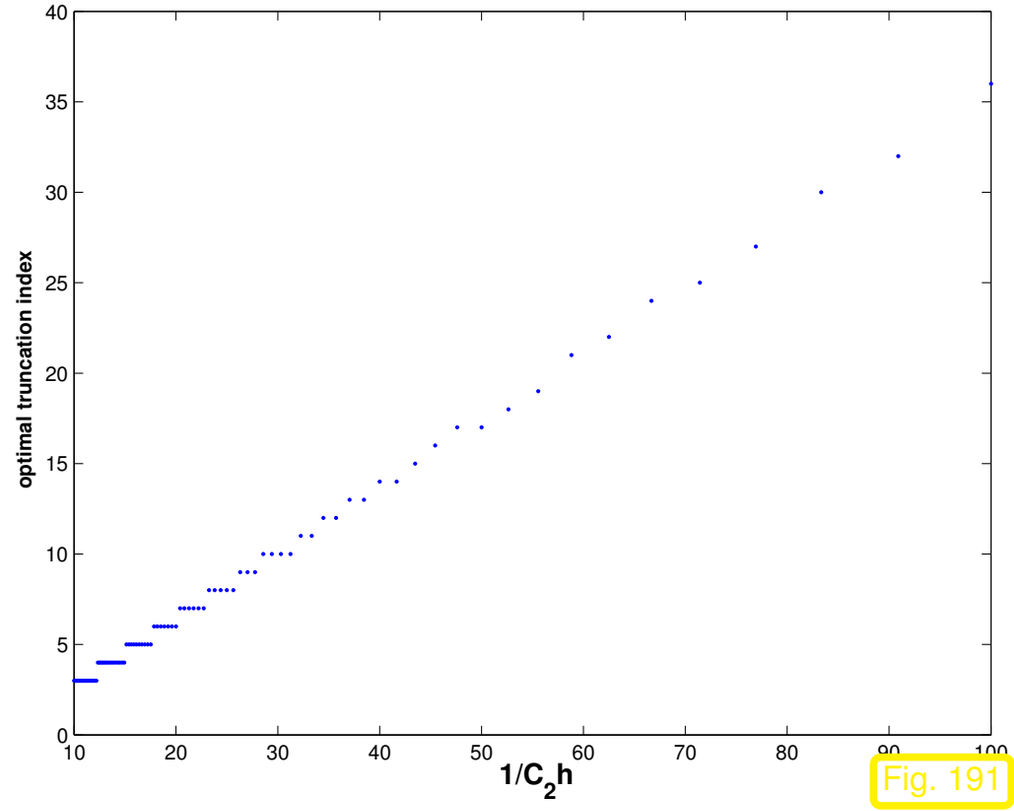


Fig. 191

P. Grohs

rev 63606, February 17, 2014

✎ Notation: $[x] \hat{=}$ integral part of $x > 0$

(4.4.80) does follow from curve sketching of $x \mapsto (ax)^x$, $x > 0$.

$$\Rightarrow \quad \left\| \mathbf{\Psi}^h \mathbf{y} - \widetilde{\mathbf{\Phi}}^h_{h,\ell_{\mathrm{opt}}} \mathbf{y} \right\| \leq C_1 h \exp(-\ell_{\mathrm{opt}}) \leq C_1 h \exp(-\gamma/h) \ , \quad \gamma := \frac{1}{C_2 e} > 0 \ . \qquad (4.4.81)$$

Schranke exponentially small for $h \to 0$, vgl. (4.4.63)

Note: (4.4.81) $\leftrightarrow$ consistency error estimate (2.1.18)

Now we derive a estimate for the difference of the numerical solution of the trajectory of the optimal cropped modified equation $\dot{\widetilde{\mathbf{y}}} = \widetilde{\mathbf{f}}_{h,\ell_{\mathrm{opt}}}(\widetilde{\mathbf{y}})$, cf. (4.4.62).

P. Grohs

rev 63606, February 17, 2014

Consider: IVP $\boxed{\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) \ \mathbf{y}(0) = \mathbf{y}_0 \in D}$ on $[0, T]$, endtime $T \in J(\mathbf{y}_0)$

Fix initial value $\mathbf{y}_0 \in D$, end time $T \in J(\mathbf{y}_0)$

compact neighborhood of the trajectory of the solution $t \mapsto \mathbf{y}(t)$, $0 \leq t \leq T$,
of the IVP $\boxed{\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}), \quad \mathbf{y}(0) = \mathbf{y}_0}$

✎ Notations: $\widetilde{\mathbf{y}} \hat{=}$ Solution of the IVP $\dot{\widetilde{\mathbf{y}}} = \widetilde{\mathbf{f}}_{h,\ell_{\mathrm{opt}}}(\widetilde{\mathbf{y}})$, $\widetilde{\mathbf{y}}(0) = \mathbf{y}_0$

(ℓ_{opt} from (4.4.80) with C_2 thm. 4.4.66 relatively to K)

$\left(\mathbf{y}_k^h\right)_k := \left((\boldsymbol{\Psi}^h)^k \mathbf{y}_0\right)_k$, $k \in \{0,\dots,[T/h]\}$: meshgrid function produced by the one-step-method with stepwidth $h > 0$ (numerical approximation solution)

Lemma 4.4.82 (Convergence of the optimal cropped modified equation)**.**

- *There is a* compact *neighborhood* $K \subset D$ *of* $\mathbf{y}_0$*, such that* $\mathbf{y}_k^h \in K$ *for all* $k \in \{0,\dots,[T/h]\}$*, if* h *is sufficiently small.*
- *We have the assumption from Thm. 4.4.66 (Analyticity assumption).*
- *The discrete evolution of the OSM does have the representation* $\boldsymbol{\Psi}^h\mathbf{y} = \mathbf{y} + h\boldsymbol{\psi}(\mathbf{y},h)$ *with a increment function uniformly lipschitz-continuous on* K $\boldsymbol{\psi}$*, e.g., cf. 2.1.24,*

$$\exists L > 0\colon \ \|\boldsymbol{\psi}(\mathbf{z},h) - \boldsymbol{\psi}(\mathbf{w},h)\| \le L\,\|\mathbf{z}-\mathbf{w}\| \quad \forall \mathbf{z},\mathbf{w} \in K,\ |h| \text{ sufficiently small.}$$

Then there is $h_0 > 0$ *and from* h_0 *independent constant* $C > 0$, $\gamma > 0$ *such that*

$$\left\|\widetilde{\mathbf{y}}(hk) - \mathbf{y}_k^h\right\| \le C(\exp(hkL) - 1)\exp(-\gamma/h) \quad \forall k \in \{0,\dots,[T/h]\}\ , \quad \forall 0 < h < h_0\ .$$

P. Grohs

rev 63606, February 17, 2014

Proof. See (4.4.62) and the remarks there:

The proof of Thm. 2.1.19 can be converted almost untouched, after (2.1.23) has been replaced by (4.4.81). See also sect. 2.1.4 for the technique of the proof. □

$T < \dfrac{\gamma}{Lh}$	$\Rightarrow$	"exponentially small" error of the one-step-method relative to optimal cropped modified equation

Next point: Is an optimal cropped equation really generated by a "small" perturbation of the underlying ODE?

Lemma 4.4.83 (Perturbation estimate for optimal cropped modified equation). $\rightarrow$ *[35, Thm. 2]*

Next to the assumptions of Thm. 4.4.66 (Analyticity assumption) there is for each compact $K \subset D$ a (sufficiently small) $h > 0$ independent constant $C > 0$ such that

$$\left\| \widetilde{\mathbf{f}}_{h,\ell}(\mathbf{y}) - \mathbf{f}(\mathbf{y}) \right\| \leq C h^p \quad \forall \mathbf{y} \in K \ , \quad \forall \ell \in \mathbb{N} \ .$$

Proof. Extension to the proof of Thm. 4.4.66, see the assumptions made there and the used notations. Demonstration for the explicit euler-method, e.g. $p = 1$.

From the definition of $\widetilde{f}_{h,\ell}$, $\rightarrow$ lemma 4.4.61,

$$\widetilde{f}_{h,\ell}(y) - f(y) = \sum_{j=1}^{\ell} h^j \Delta f_j(y) \ .$$

Idea: Use estimation of the modificator functions Δf_j from the proof of Thm. 4.4.66

concrete: from (4.4.73) with $\alpha = 0$

$$|\widetilde{f}_{h,\ell}(y) - f(y)| \leq |h| \left(\frac{2M}{R} + bM \sum_{j=2}^{\ell} |h|^{j-1} \left(\frac{jcM}{R} \right)^j \right) \ .$$

$$|h| \leq \frac{R}{(\ell+1)2cM} \quad \Rightarrow \quad |\widetilde{f}_{h,\ell}(y) - f(y)| \leq |h| \Big(\frac{2M}{R} + \frac{2bcM^2}{R} \underbrace{\sum_{j=2}^{\ell} 2^{-j} j \left(\frac{j+1}{\ell+1} \right)^j}_{\text{bounded}} \Big) \; .$$

sum

$$\ell \mapsto \sum_{j=2}^{\ell} 2^{-j} j \left(\frac{j+1}{\ell+1} \right)^j \qquad (4.4.84)$$

▷

MATLAB: Numcourses/NumerischeMathematik/-MatlabRevised/jsum

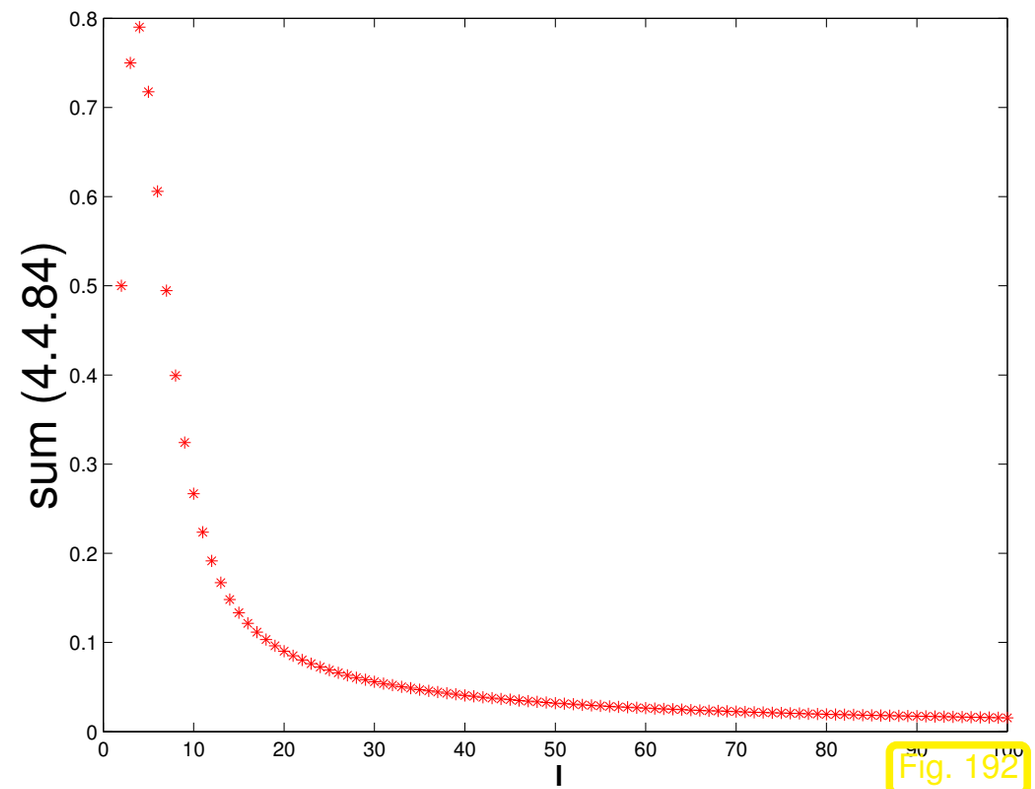


Fig. 192

P. Grohs

rev 63606, February 17, 2014

Hence: $|\widetilde{f}_{h,\ell}(y) - f(y)| \leq Ch$ for small h, $\forall y \in K$, with $C > 0$ independent of ℓ. $\square$

The vectorfield of the optimal cropped modified equation is "$O(h^p)$-close" to $\mathbf{f}$

Remark 4.4.85 (Stepwidth condition for "longtime integration").

- Consider IVP $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$, $\mathbf{y}(0) = \mathbf{y}_0$, on $[0,T] \subset J(\mathbf{y}_0)$, $\mathbf{f}$ holomorphic.
- ESV $\mathbf{y}_1 = \mathbf{\Psi}^h \mathbf{y}_0$ the consistency order $p \in \mathbb{N}$.

stepwidth for exact numerical solution ($\|\mathbf{y}(hk) - \mathbf{y}_k\|$ small) on $[0,T]$

$$\text{Thm. 2.1.19} \quad \Rightarrow \quad h^p \exp(LT) \ll 1 \quad \Rightarrow \quad \boxed{h = O(\exp(-T/p))} \; .$$

Stepwidth condition for acceptable $(*)$ numerical solution ($\|\widetilde{\mathbf{y}}(hk) - \mathbf{y}_k\|$ small) on $[0,T]$

$$\text{lemmas 4.4.82, 4.4.83} \quad \Rightarrow \quad h < \frac{\gamma}{LT} \quad \Rightarrow \quad \boxed{h = O(T^{-1})} \; .$$

$(*)$ "generic acceptable" relative to general additive perturbation of $\mathbf{f}$. (More strict: structure-preserving acceptable, see beginning of Sect. 4.4.3) $\triangle$

Theory: Backward analysis $\rightarrow$ [21, Ch. IX], [31], [**?**]

Hamiltonian ODE $\dot{\mathbf{y}} = \mathbf{J}\,\mathbf{grad}\,H(\mathbf{y})$, H analytic (holomorphic) in D, $\{\mathbf{y}_k\}_k$ produced by symplectic ones-step-method *with* <u>*uniform*</u> *time-stepwidth* $h > 0$.

$\Downarrow$

$\|\mathbf{y}_k - \widetilde{\mathbf{y}}(kh)\| \leq C\exp(-\gamma h^{-1})$ mit $C, \gamma > 0$ independent of h,
for "exponentially long times", $t \mapsto \widetilde{\mathbf{y}}(t) \mathrel{\hat{=}}$ *Solution of a hamiltonian differential equation.* $\frac{d}{dt}\widetilde{\mathbf{y}} = \mathbf{J}^{-1}\,\mathbf{grad}\,\widetilde{H}_h(\mathbf{y})$, $\widetilde{\mathbf{y}}(0) = \mathbf{y}(0)$, $\widetilde{H}_h \mathrel{\hat{=}}$ *"slightly perturbed" Hamilton-function*: $H - \widetilde{H}_h = O(h^p)$.

4.4.5 Structure-preserving modified equations

According to Bem. 4.4.43 we have to show that the vectorfields $\widetilde{\mathbf{f}}_{h,\ell}$ of the cropped modified equations structural properties ($\mathbf{f} \in V$ of Rem. 4.4.14) of $\mathbf{f}$ inherit. Focus is on hamiltonian differential equations ($\rightarrow$ Def. 1.2.20) $\leftrightarrow$ symplectizity ($\rightarrow$ Def. 4.4.12)

Example 4.4.86 (modified equation for symplectic euler-method). $\rightarrow$ [31, Sect. 5.1.2]

separated hamilton-function with potential $U : \mathbb{R}^n \mapsto \mathbb{R}$, cf. (1.2.28)

$$H(\mathbf{p},\mathbf{q}) := \tfrac{1}{2}\left\|\mathbf{p}\right\|^2 + U(\mathbf{q}) \;, \quad \mathbf{p},\mathbf{q} \in \mathbb{R}^n \;.$$

► hamiltonian differential equation:

$$\dot{\mathbf{p}} = -\,\mathbf{grad}\,U(\mathbf{q}) \quad , \quad \dot{\mathbf{q}} = \mathbf{p} \;. \tag{4.4.87}$$

① explicite euler-method for (4.4.87), stepwidth $h > 0$ (Consistency order 1):

$$\mathbf{p}_1 = \mathbf{p}_0 - h\,\mathbf{grad}\,U(\mathbf{q}_0) \quad , \quad \mathbf{q}_1 = \mathbf{q}_0 + h\mathbf{p}_0 \;.$$

Taylor expansion & (4.4.87) & (4.4.54) ➢ first modificator function

② symplectic euler-method (4.4.26), stepwidth $h > 0$ (consistency order 1):

$$\mathbf{p}_1 = \mathbf{p}_0 - h\,\mathbf{grad}\,U(\mathbf{q}_1) \quad , \quad \mathbf{q}_1 = \mathbf{q}_0 + h\mathbf{p}_0 \;.$$

taylor expansion & (4.4.87) & (4.4.54) ➢ modificator function

Here: modified equation of second order is hamiltonian differential equation **!**

concrete: mathematical pendulum $\longrightarrow$ Ex. 1.2.17

$$n = 1 \;, \quad H(p,q) = \tfrac{1}{2}p^2 - \cos q \;.$$

"exact" trajectories $\triangleright$

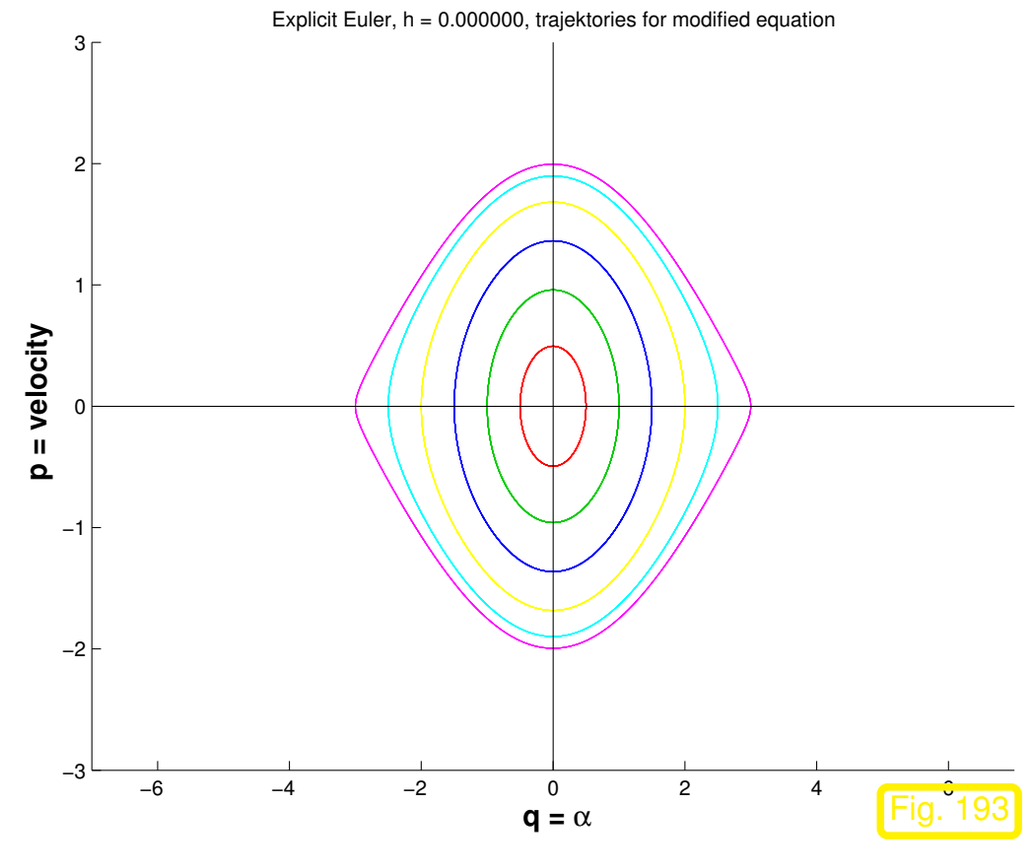
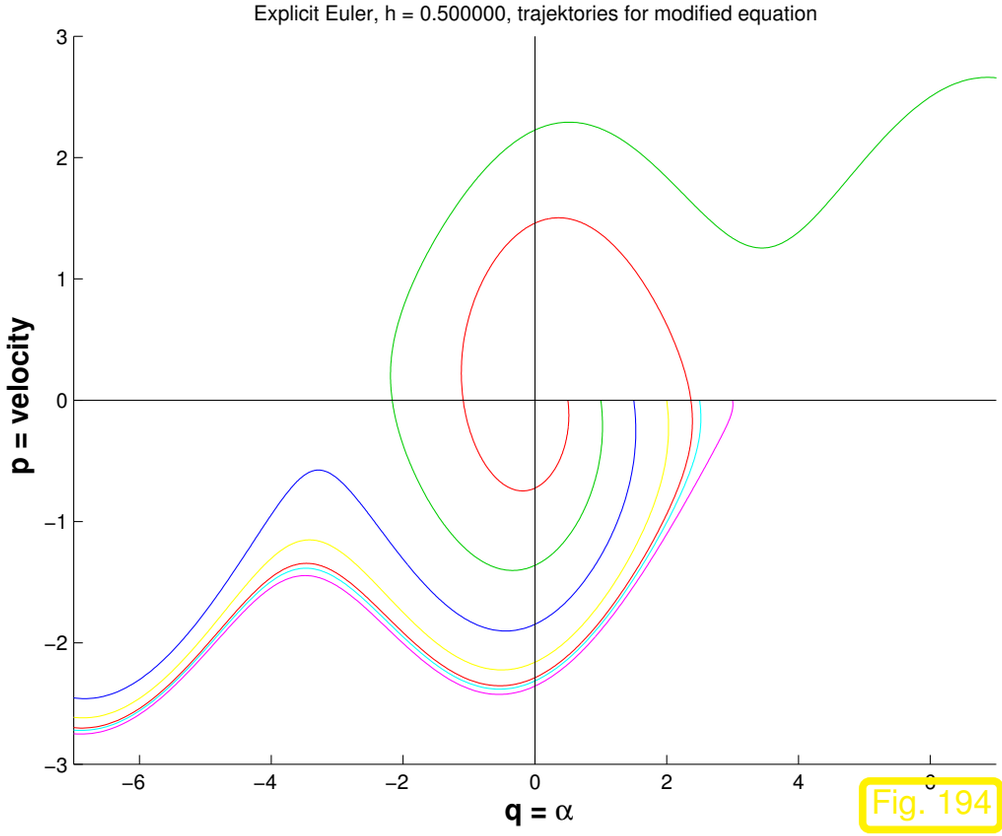


Fig. 193

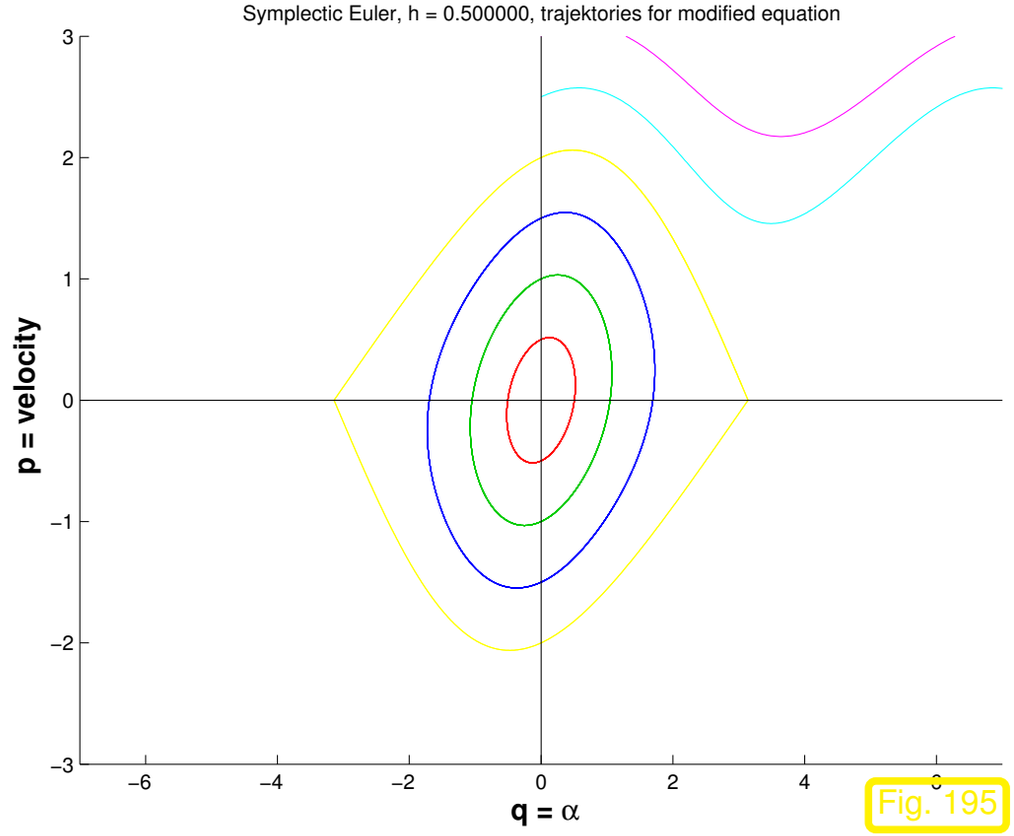
trajectories to modified equation of order 2

P. Grohs

rev 63606, February 17, 2014



Expl. Euler, $h = 0.5$



symplectical Euler, $h = 0.5$ ◇

Theorem 4.4.88 (Symplecticity of the modified evolutions)**.**
Let $\mathbf{\Psi}^h$ be the discrete evolution of a symplectical one-step method ($\rightarrow$ Def. 4.4.18) for the hamiltonian differential equation $\dot{\mathbf{y}} = \mathbf{J}^{-1} \cdot \mathbf{grad}\, H(\mathbf{y})$ with smooth hamiltonian function $H : D \subset \mathbb{R}^{2n} \mapsto \mathbb{R}$, D starshaped.
Then the cropped modified equation $\dot{\widetilde{\mathbf{y}}} = \widetilde{\mathbf{f}}_{h,\ell}(\widetilde{\mathbf{y}})$ from Lemma 4.4.61 are also hamiltonian for all $\ell \in \mathbb{N}$ and all (sufficiently small) $h > 0$.

Proof. (of Thm. 4.4.88) Idea: Induction on ℓ

Induction basis: For $\ell \leq p$: $\widetilde{\mathbf{f}}_{h,\ell}(\mathbf{y}) = \mathbf{f}(\mathbf{y}) = \mathbf{J}^{-1} \cdot \mathbf{grad}\, H(\mathbf{y})$ ✔

$\widetilde{\mathbf{\Phi}}^t \hat{=}$ evolution operator to $\dot{\widetilde{\mathbf{y}}} = \widetilde{\mathbf{f}}_{h,\ell}(\widetilde{\mathbf{y}}) = \mathbf{J}^{-1}\, \mathbf{grad}\, \widetilde{H}_\ell(\mathbf{y})$ (induction hypothesis !)

$\Rightarrow$ For fixed t: $\widetilde{\mathbf{\Phi}}^t : D \mapsto \mathbb{R}^{2n}$ ist symplectic ($\rightarrow$ Def. 4.4.12)

After (4.4.52) & (4.4.54), Lemma 4.4.61 for $h \rightarrow 0$

$$\widetilde{\mathbf{\Phi}}^h \mathbf{y}_0 - \mathbf{\Psi}^h \mathbf{y}_0 = -\Delta \mathbf{f}_{\ell+1}(\mathbf{y}_0) h^{\ell+2} + O(h^{\ell+3}) \quad \forall \mathbf{y}_0 \,. \tag{4.4.89}$$

$$(D_{\mathbf{y}}\widetilde{\mathbf{\Phi}}^h)(\mathbf{y}_0) - (D_{\mathbf{y}}\mathbf{\Psi}^h)(\mathbf{y}_0) = -(D_{\mathbf{y}}\Delta\mathbf{f}_{\ell+1})(\mathbf{y})h^{\ell+2} + O(h^{\ell+3}) \; . \tag{4.4.90}$$

$\widetilde{\mathbf{\Phi}}^h$, $\mathbf{\Psi}^h$ = symplectic maps ($\rightarrow$ Def. 4.4.12, argument $\mathbf{y}_0$ omited) $\Rightarrow$

$$\underbrace{(D_{\mathbf{y}}\widetilde{\mathbf{\Phi}}^h)^T\mathbf{J}D_{\mathbf{y}}\widetilde{\mathbf{\Phi}}^h}_{=\mathbf{J}} = \underbrace{(D_{\mathbf{y}}\mathbf{\Psi}^h)^T\mathbf{J}D_{\mathbf{y}}\mathbf{\Psi}^h}_{=\mathbf{J}}$$
$$+ h^{\ell+2}\Big((D_{\mathbf{y}}\mathbf{\Psi}^h)^T\mathbf{J}D_{\mathbf{y}}\Delta\mathbf{f}_{\ell+1} + (D_{\mathbf{y}}\Delta\mathbf{f}_{\ell+1})^T\mathbf{J}D_{\mathbf{y}}\mathbf{\Psi}^h\Big) + O(h^{\ell+3}) \; .$$
$$\Rightarrow \quad 0 = (D_{\mathbf{y}}\mathbf{\Psi}^h)^T\mathbf{J}D_{\mathbf{y}}\Delta\mathbf{f}_{\ell+1} + (D_{\mathbf{y}}\Delta\mathbf{f}_{\ell+1})^T\mathbf{J}D_{\mathbf{y}}\mathbf{\Psi}^h + O(h) \; .$$

For consistent OSM, see Lemma 2.1.9: $D_{\mathbf{y}}\mathbf{\Psi}^h = \mathbf{I} + O(h)$ for $h \rightarrow 0$.

$$\overset{h\to 0}{\Rightarrow} \quad 0 = \mathbf{J}D_{\mathbf{y}}\Delta\mathbf{f}_{\ell+1} + (D_{\mathbf{y}}\Delta\mathbf{f}_{\ell+1})^T\mathbf{J} \quad \Rightarrow \quad D_{\mathbf{y}}(\mathbf{J}\Delta\mathbf{f}_{\ell+1}) = (D_{\mathbf{y}}(\mathbf{J}\Delta\mathbf{f}_{\ell+1}))^T \; .$$

▶ Application of Lemma 4.4.17 (Integrability lemma) on $\mathbf{J}\Delta\mathbf{f}_{\ell+1}$. □

P. Grohs

rev 63606, February 17, 2014

symplectic integrators yield structure preserving **acceptable** $(*)$ discrete evolutions for (smooth) conservative mechanical systems.

(exponentially exact) solution of an evolution with "slightly perturbed" (namely $O(h^p)$, see Lemma 4.4.83 hamiltonian function ➢ Backward analysis $\rightarrow$ Sect. 4.4.3

Explanation of "longtime energy conservation" of symplectical integrators by backward analysis:

☞ solution of the OSM (Consistency order p) is "exponentially exact" ($\rightarrow$ Lemma 4.4.82) approximation of the solution of a (optimal cropped) modified equation

This is a hamiltonian differential equation ($\rightarrow$ Def. 1.2.20) with a (rel. to H) around $O(h^p)$ perturbed hamiltonian-function $\widetilde{H}(\mathbf{y})$ ($\rightarrow$ Thm. 4.4.83).

Details:

- symplectical OSM of the consistency order p
- stepwidth $h < h^*$ $\Rightarrow$ numerical solution $(\mathbf{y}_k)_k \subset K \subset D$, K *compact* subset $K \subset D$ of the state space.
- K is starshaped (not necessary [21, Sect. XI.3.2])
- $\mathbf{f}$ satisfies analyticity condition rel. to K

$(t, \mathbf{y}) \mapsto \widetilde{\mathbf{\Phi}}_h^t \mathbf{y} \hat{=}$ evolution operator to the (optimal cropped) modified equation $\dot{\widetilde{\mathbf{y}}} = \widetilde{\mathbf{f}}_{h,\ell_{\text{opt}}}(\widetilde{\mathbf{y}})$.

Estimation (4.4.81) $\Rightarrow$ $\left\| \mathbf{\Psi}^h \mathbf{y} - \widetilde{\mathbf{\Phi}}_h^h \mathbf{y} \right\| \leq Ch \exp(-\gamma/h) \quad \forall \mathbf{y} \in K, \forall h < h^*$.

Thm. 4.4.88 $\Rightarrow$ $\widetilde{\mathbf{f}}_{h,\ell_{\mathrm{opt}}}(\mathbf{y}) = \mathbf{J}^{-1}\,\mathbf{grad}\,\widetilde{H}_h(\mathbf{y})$ mit $\widetilde{H}_h : K \mapsto \mathbb{R}$ holomorphic .

$$\widetilde{H}_h : K \mapsto \mathbb{R} \quad \text{holomorphic} \quad \overset{K\text{ compact}}{\Rightarrow} \exists L > 0\colon \quad |\widetilde{H}_h(\mathbf{y}) - \widetilde{H}_h(\mathbf{z})| \leq L\,\|\mathbf{y}-\mathbf{z}\| \quad \forall \mathbf{y},\mathbf{z} \in K\ .$$

"telescopic sum argument": da $\widetilde{H}_h(\widetilde{\mathbf{\Phi}}_h^t\mathbf{y}) = \widetilde{H}_h(\mathbf{y})$ for all $t \in J(\mathbf{y})$, $\mathbf{y} \in K$ (hamiltonian-function is invariant of a hamiltonian ODE, see Lemma 1.2.23)

$$\begin{aligned}|\widetilde{H}_h(\mathbf{y}_k) - \widetilde{H}_h(\mathbf{y}_0)| &\leq \sum_{j=0}^{k-1} |\widetilde{H}_h(\mathbf{y}_{j+1}) - \widetilde{H}_h(\mathbf{y}_j)| = \sum_{j=0}^{k-1} |\widetilde{H}_h(\mathbf{\Psi}^h\mathbf{y}_j) - \widetilde{H}_h(\widetilde{\mathbf{\Phi}}_h^h\mathbf{y}_j)| \\ &\leq \sum_{j=0}^{k-1} L \left\| \mathbf{\Psi}^h\mathbf{y}_j - \widetilde{\mathbf{\Phi}}_h^h\mathbf{y}_j \right\| \leq CL \sum_{j=0}^{k-1} h\exp(-\gamma/h) \leq CLhk\exp(-\gamma/h)\ .\end{aligned}$$

Thm. 4.4.88 $\Rightarrow$ $\exists C > 0\colon \quad \max\limits_{\mathbf{y}\in K} |\widetilde{H}_h(\mathbf{y}) - H(\mathbf{y})| \leq Ch^p \quad \forall h < h^*$.

$$\Rightarrow \quad \boxed{|H(\mathbf{y}_k) - H(\mathbf{y}_0)| \leq C(Lhk\exp(-\gamma/h) + h^p) \quad \forall h < h^*}\ .$$

➢ "energy error" of the numerical solution of size $O(h^p)$ for "exponentially long time"

☞ This is the explanation for the conjecture from Ex. (4.4.34) !

Theorem 4.4.91 (Longtime conservation of energy with symplectic integration)**.**
For the hamiltonian ODE $\dot{\mathbf{y}} = \mathbf{J}^{-1}\,\mathbf{grad}\,H(\mathbf{y})$ *(→ Def. 1.2.20) and an order* p *consistent symplectic one-step-method (→ Def. 4.4.18) the conditions of Thm. 4.4.66 are satisfied.*
For sufficiently small (uniform !) stepwidth h *we have* $(\mathbf{\Psi}^h)^k\mathbf{y} \in K$ *for all* $k \in \mathbb{N}_0$ *and* $\mathbf{y} \in K_0$*, where* $K, K_0 \subset D$ *compact. Then there is* $C > 0$ *with*

$$|H((\mathbf{\Psi}^h)^k\mathbf{y}_0) - H(\mathbf{y}_0)| \leq C(hk\exp(-\gamma/h) + h^p) \quad \forall h \text{ sufficiently small, } \forall \mathbf{y}_0 \in K_0 \,.$$

$$\boxed{T \lesssim \exp(O(h^{-1})) \quad \Longrightarrow \quad H(\mathbf{y}_k) - H(\mathbf{y}_0) = O(h^p)}$$

P. Grohs

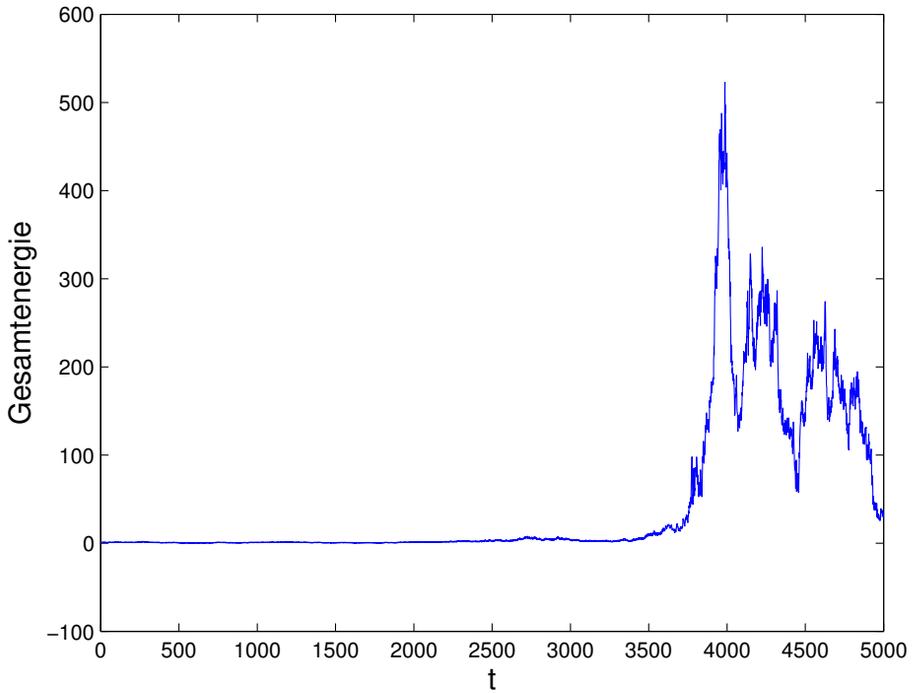
rev 63606, February 17, 2014

Sect. 4.4.3": Method of backward analysis requires *uniform time stepwidth*.

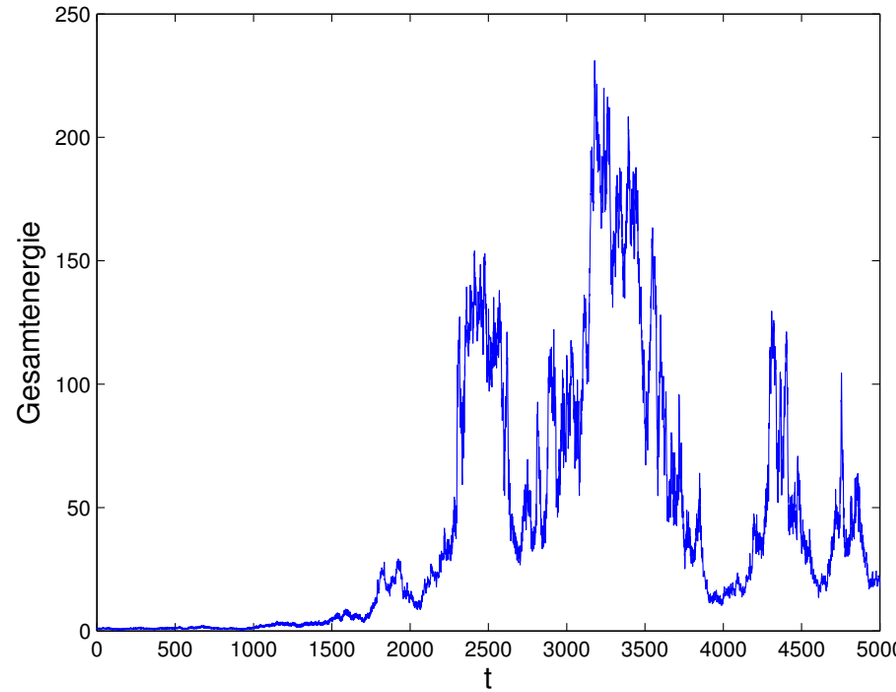
A merely theoretical constraint **?**

Example 4.4.92 (Symplectic integrators and variable stepwidth)*.* Continuation Ex. 4.4.33

Symplectic euler-method (4.4.26) for (4.4.4) on $[0,T]$, $T = 5000$. Erratic variable stepwidth $h_i = 0.5(1+0.5(\mathrm{rand}()-0.5))$, $i = 1,\dots,10000$, $p(0) = 0$, $q(0) = 7\pi/6$



Energy drift at variable stepwidth (method (4.4.26), left)



Energy drift at variable stepwidth (method (4.4.26), right)

◇

4.5 Methods for oscillatori differential equation [28]

prototype: $$\ddot{y} = -\omega^2 y \quad , \quad y(0) = y_0,\ \dot{y}(0) = v_0$$

$$\blacktriangleright \qquad y(t) = \alpha\cos(\omega t) + \beta\sin(\omega t)\ , \quad \alpha,\beta \in \mathbb{R}$$

☞ generalization (scalar): $$\ddot{y} = -\omega^2 y + g(y) \quad , \quad y(0) = y_0,\ \dot{y}(0) = v_0\ , \tag{4.5.1}$$

with Lipschitz-continuous perturbation $g : \mathbb{R} \mapsto \mathbb{R}$.

Generalization (vectorial) $$\ddot{\mathbf{y}} = -\mathbf{A}\mathbf{y} + g(\mathbf{y}) \quad , \quad \mathbf{y}(0) = \mathbf{y}_0,\ \dot{\mathbf{y}}(0) = \mathbf{v}_0\ , \tag{4.5.2}$$

$\mathbf{A} \in \mathbb{R}^{d,d}$ symmetric positive definite, $g : \mathbb{R}^d \mapsto \mathbb{R}^d$

Remark 4.5.3.

$$(4.5.1) \quad \overset{v:=\dot{y}}{\Longleftrightarrow} \quad \frac{d}{dt}\begin{pmatrix} y \\ v \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix}\begin{pmatrix} y \\ v \end{pmatrix} + \begin{pmatrix} 0 \\ g(y) \end{pmatrix}\ . \tag{4.5.4}$$

solution of (4.5.4) by variation of constants:

$$\begin{pmatrix} y(t) \\ v(t) \end{pmatrix} = \begin{pmatrix} \cos t\omega & \omega^{-1}\sin t\omega \\ -\omega\sin t\omega & \cos t\omega \end{pmatrix}\begin{pmatrix} y_0 \\ v_0 \end{pmatrix} + \int_0^t \begin{pmatrix} \omega^{-1}\sin(t-s)\omega \\ \cos(t-s)\omega \end{pmatrix} g(y(s))\,\mathrm{d}s \tag{4.5.5}$$

Remark 4.5.6. $y(t)$ solves (4.5.1) & $G' = g$ ➡ $\frac{1}{2}|\dot{y}|^2 + \frac{1}{2}\omega^2 y(t) - G(y(t)) \equiv \text{const}$.

"'energy" for ODE (4.5.1)

△

Example 4.5.7 (Standard integrators for oscillatori differential equation).

adaptive explicite RK-OSM (Sect. 2.6 for (4.5.1):

```
y0=[1;0]; f=@(t,x) [0,1;-omega^2,0]*x + 20*[0;sin(x(1))];
options=odeset('reltol',1.0e-2,'abstol',1.0e-5);
[t,y]=ode45(f,[0,1],y0,options);
```
("'energy" $\hat{=}$ invariant from Rem. 4.5.6)

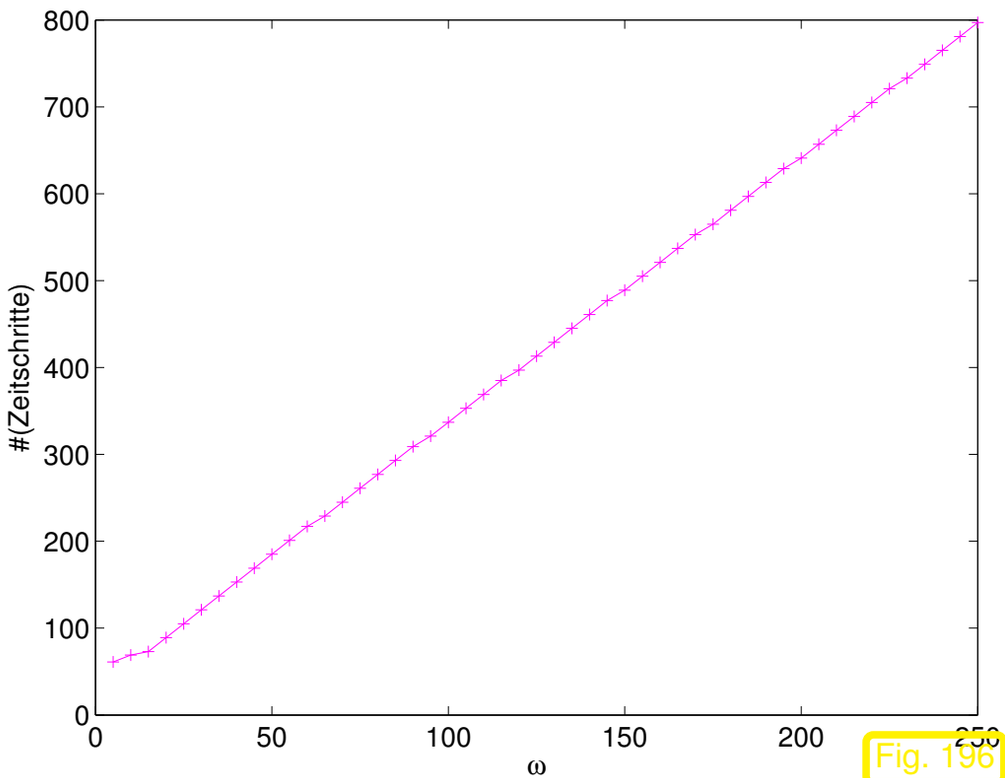


Fig. 196

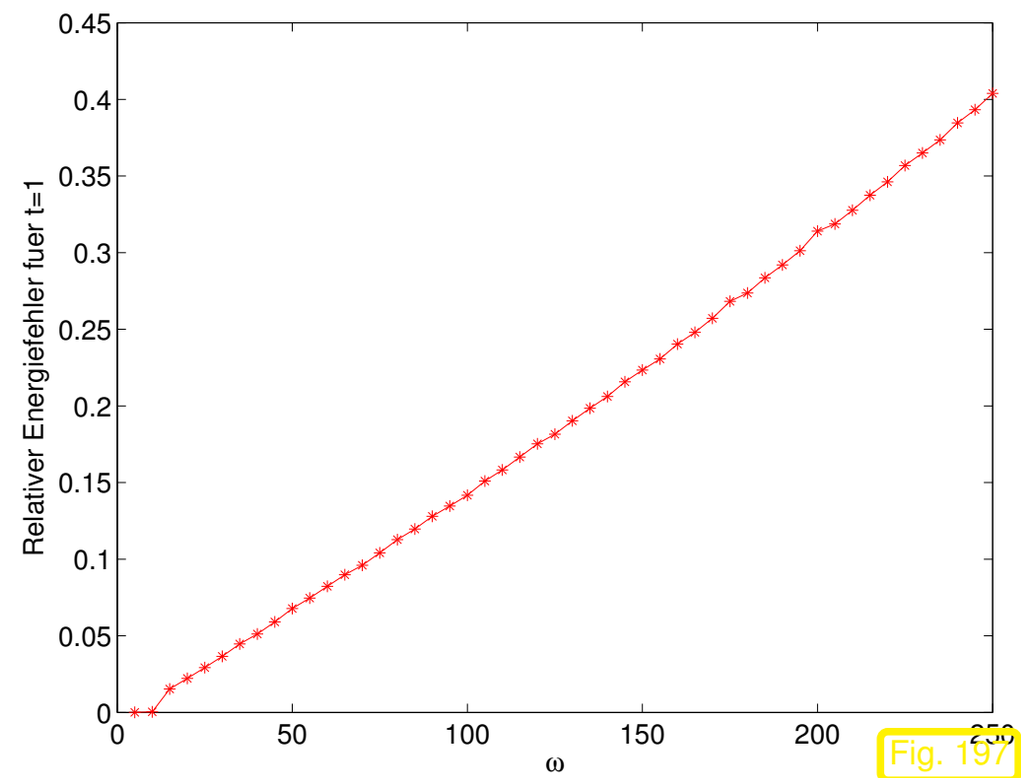


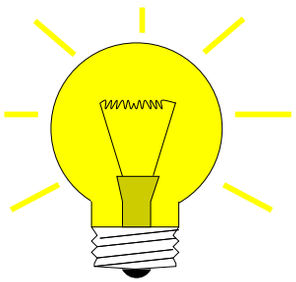
Fig. 197

$\omega \uparrow \quad \Rightarrow \quad$ Oscillations in $y(t) \uparrow \quad \Rightarrow \quad$ Number of time steps $\uparrow$

$\diamond$

Ziel: *Efficient* numerical integration of (4.5.1)/(4.5.2) also for $\omega \gg 1$ res. $\lambda_{\max}(\mathbf{A}) \gg 1$

Idea: (as for exponential integrators, see Sect. 3.7)



Use analytical representation of the solution (4.5.5) for the numerical integration:

$$\boxed{y(t \pm h) = \cos(h\omega)y(t) \pm \frac{\sin h\omega}{\omega}\dot{y}(t) + \int_0^{\pm h} \frac{\sin(\pm h - s)\omega}{\omega} \cdot g(y(t+s))\,\mathrm{d}s} \tag{4.5.8}$$

$g \equiv$ const. , (4.5.8) $\blacktriangleright$

$$y(t+h) - 2\cos(h\omega)y(t) + y(t-h) = h^2 \left(\frac{\sin(\frac{1}{2}h\omega)}{\frac{1}{2}h\omega}\right)^2 g\,. \tag{4.5.9}$$

➢ Gautschis two-step-method ($y_h(t+h)$ aus $y_h(t), y_h(t-h)$) for (4.5.1)

$$y_h(t+h) - 2\cos(h\omega)y_h(t) + y_h(t-h) = h^2 \left(\frac{\sin(\frac{1}{2}h\omega)}{\frac{1}{2}h\omega}\right)^2 g(y_h(t))\,. \tag{4.5.10}$$

Necessary: First step from (4.5.5)

$$y_h(h) = \cos(h\omega)y_0 + \frac{\sin h\omega}{\omega}v_0 + \tfrac{1}{2}h^2\left(\frac{\sin(\frac{1}{2}h\omega)}{\frac{1}{2}h\omega}\right)^2 g(y_0)\ . \tag{4.5.11}$$

Approximation of derivation: From (4.5.8) for $g \equiv \mathrm{const}$:

$$\blacktriangleright\quad y(t+h) - y(t-h) = 2h\,\frac{\sin h\omega}{h\omega}\dot{y}(t) \quad\Rightarrow\quad v_h(t) = \frac{h\omega}{\sin h\omega}\cdot\frac{y_h(t+h) - y_h(t-h)}{2h}\ . \tag{4.5.12}$$

Remark 4.5.13. Gautschi-method (4.5.10), (4.5.11) for vectorial problem (4.5.2) **?**

Replace $\cos(h\omega) \mapsto \cos h\mathbf{A}$, $\left(\frac{\sin(\frac{1}{2}h\omega)}{\frac{1}{2}h\omega}\right)^2 \mapsto 4(h\mathbf{A})^{-2}\sin^2(\frac{1}{2}h\mathbf{A})$.

△

Example 4.5.14 (Gautschis two-step-method).

Initial value problem from type (4.5.1) on $[0,1]$:

$$\ddot{y} = -\omega^2 y + \sin y \quad , \quad y(0) = 1\ , \quad \dot{y}(0) = 0\ .$$

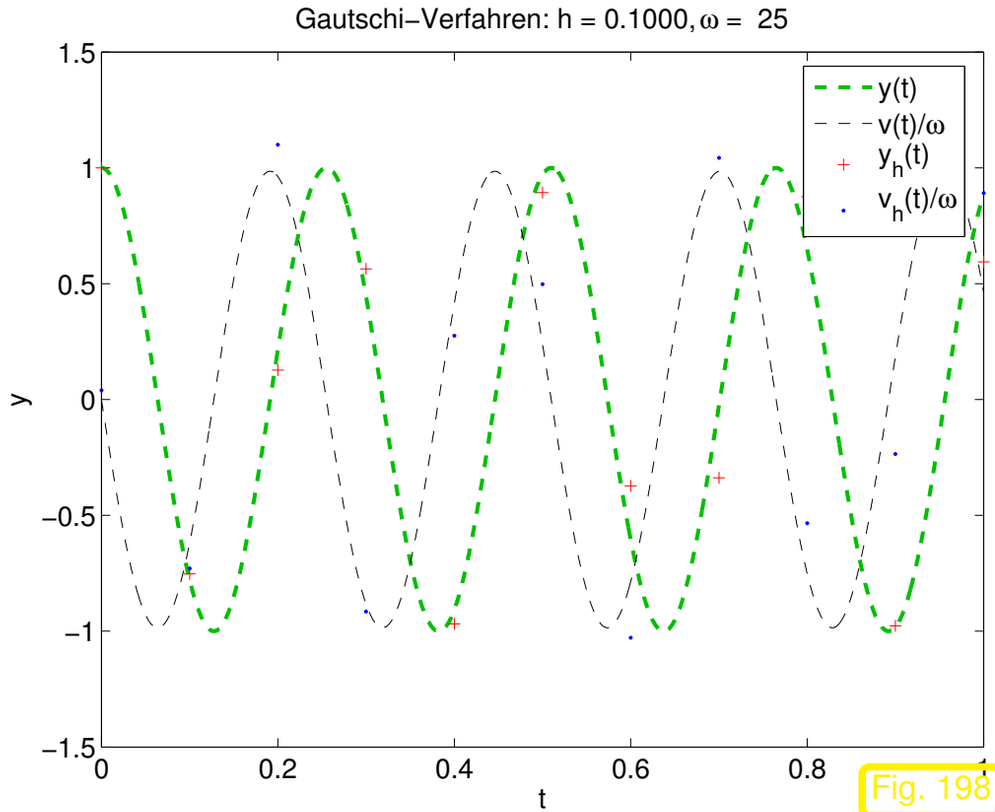


Fig. 198

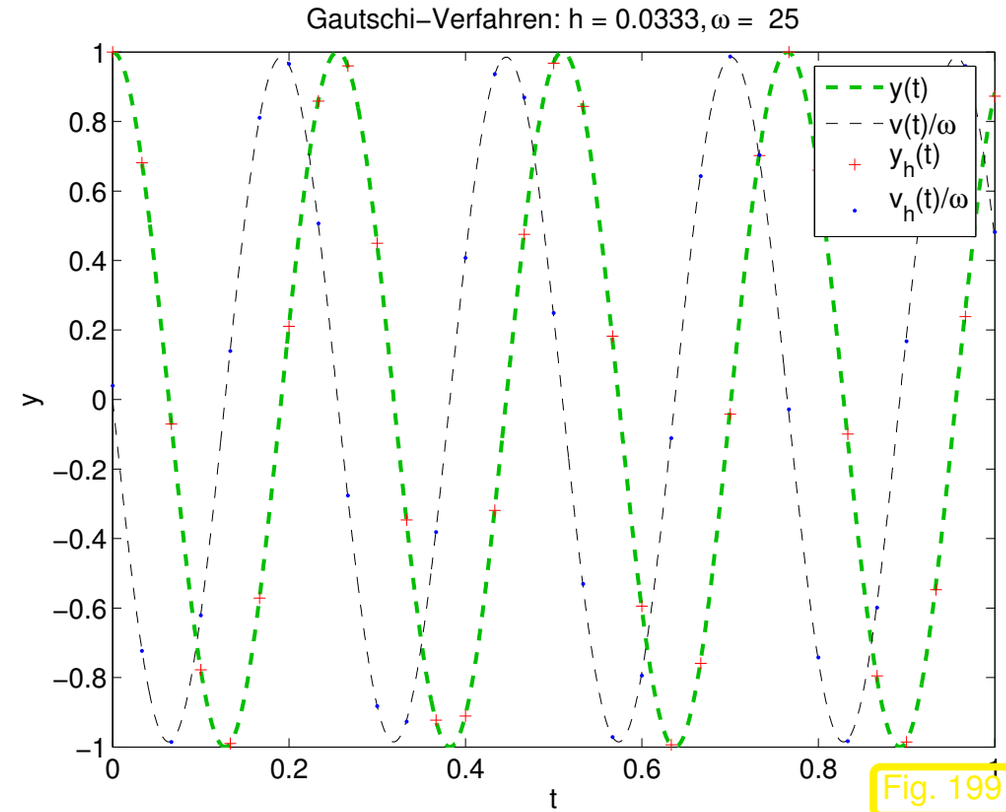


Fig. 199

$\omega = 25$: y_h follows (oscillatori solution), also if $h \approx \frac{2\pi}{\omega}$

relative error in "‘energy” ($\rightarrow$ Rem. 4.5.6) for $t = 1$:

P. Grohs

rev 63606, February 17, 2014

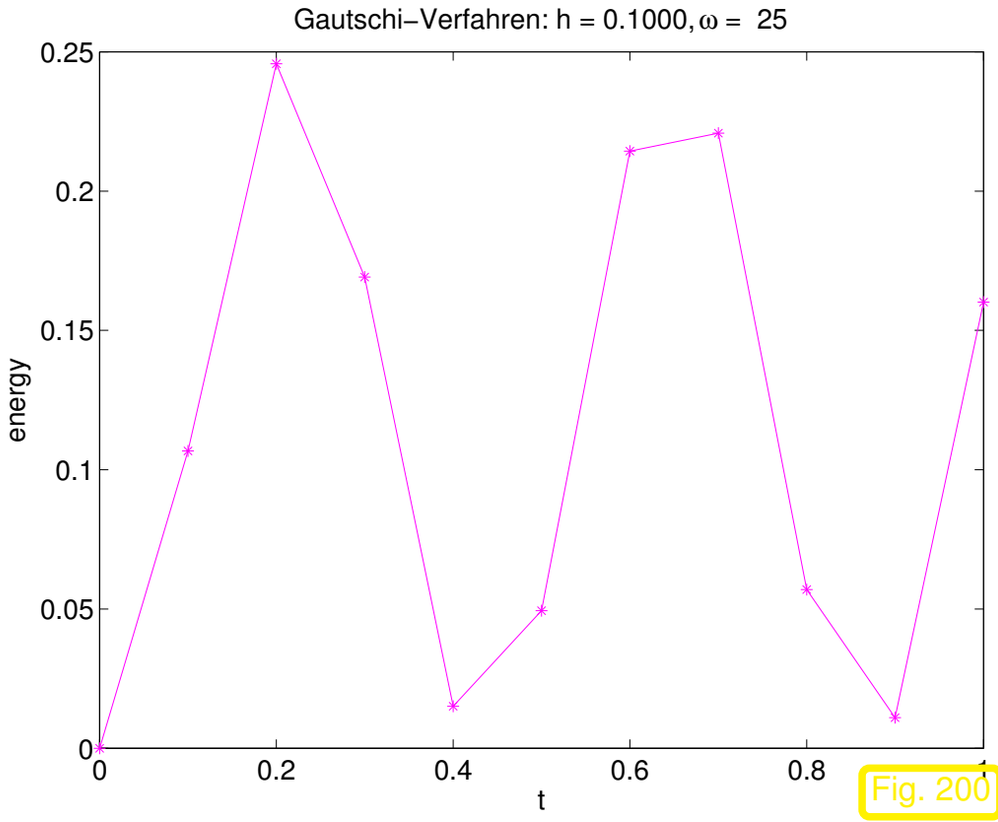


Fig. 200

Zeitschritt $h = 0.1$

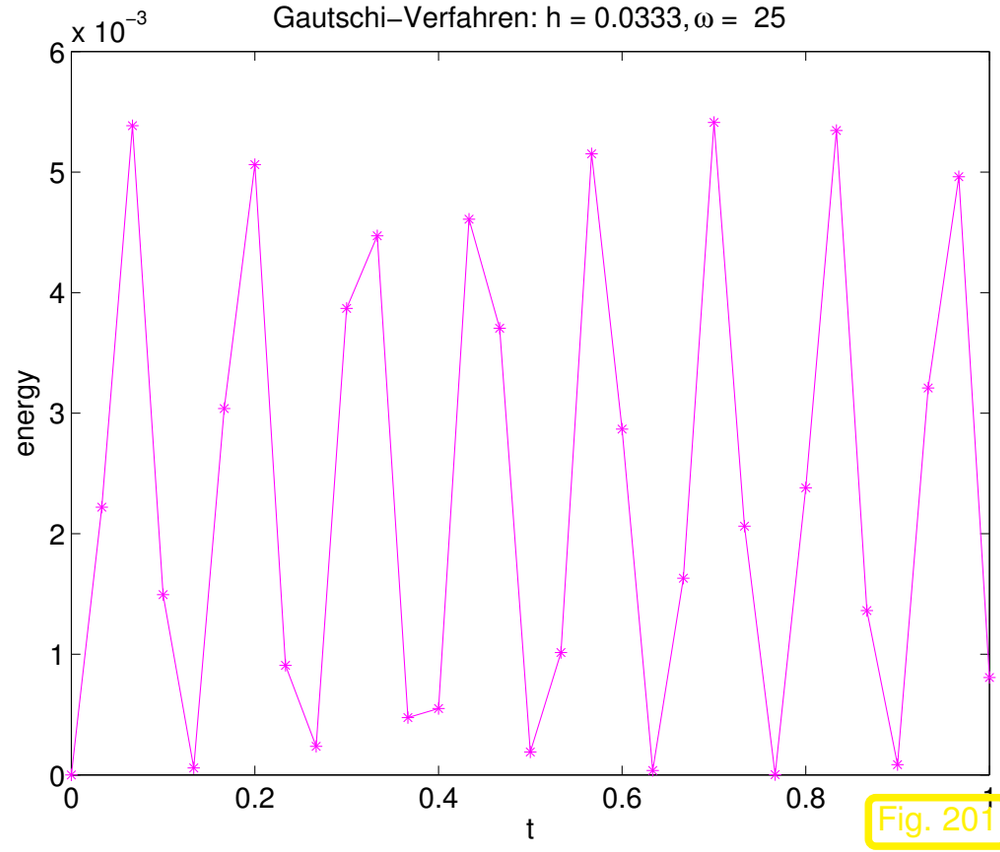


Fig. 201

time step $h = 0.033$

P. Grohs

rev 63606, February 17, 2014

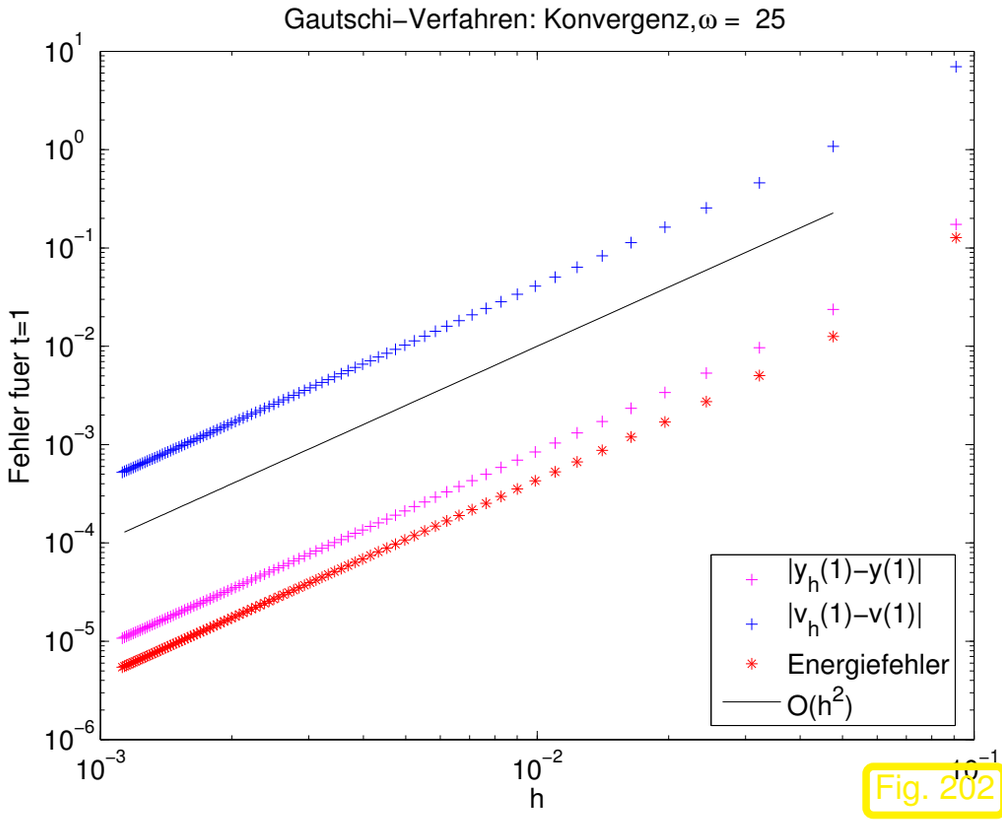


Fig. 202

Observation:

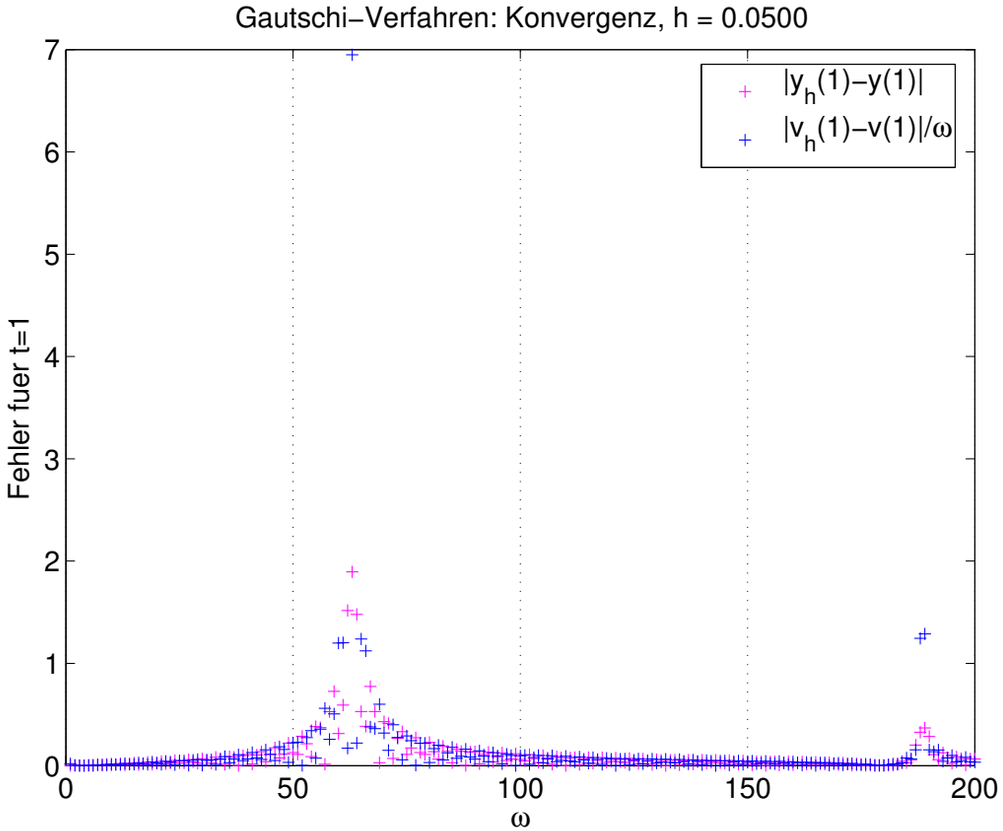
All errors $\approx O(h^2)$

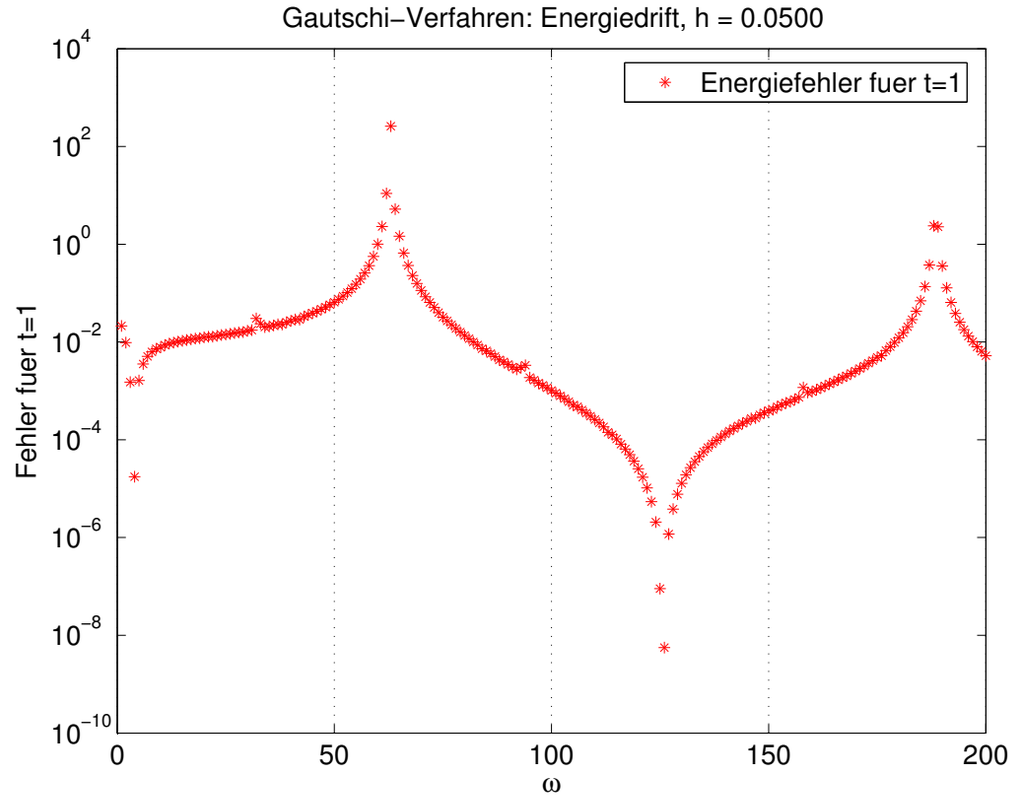
$\Updownarrow$

Convergence order 2

Goal reached **?**

Error for fix h in dependence of ω:



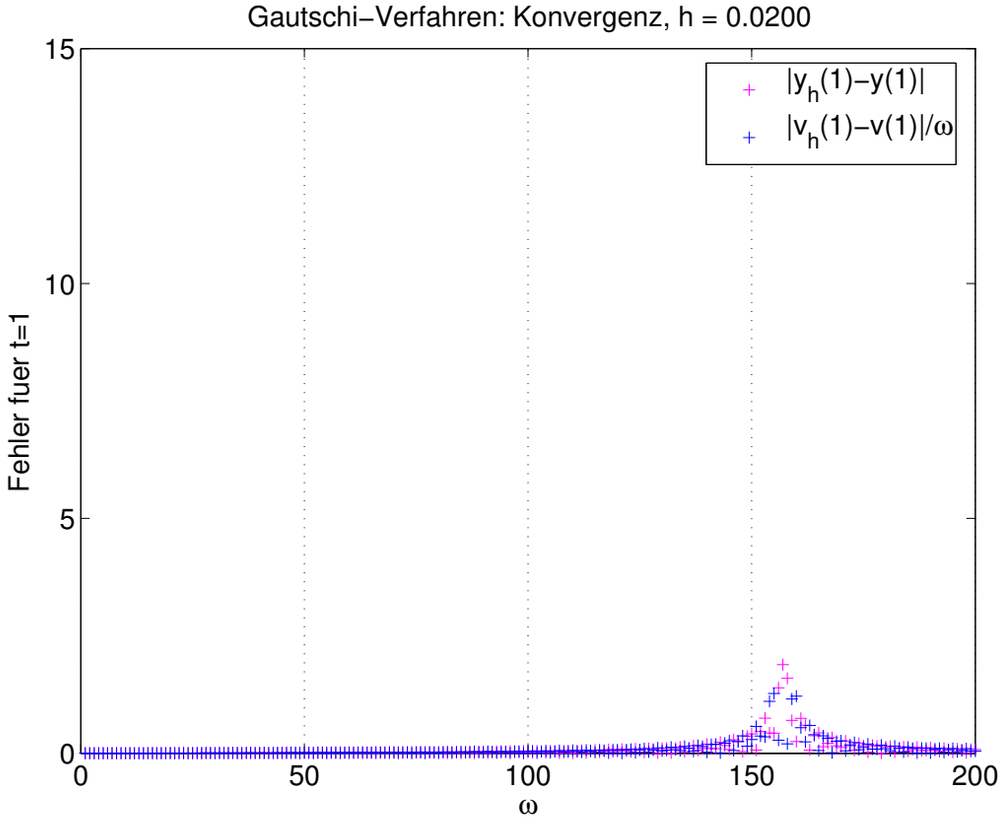


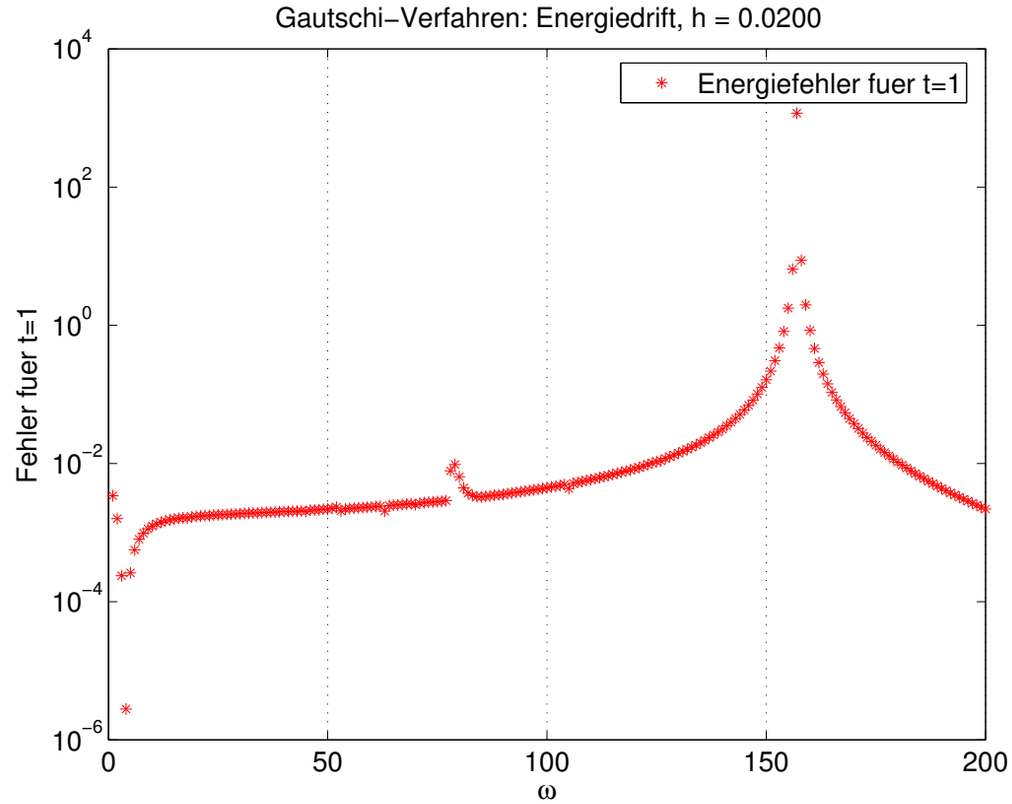
$h = 0.05$: Was pasiert für $\omega \approx 61$, $\omega \approx 123$, $\omega \approx 185$ **?**

instability ?

P. Grohs

rev 63606, February 17, 2014





➢ h-dependence on critical frequencies

h-dependence on critical frequencies

Logarithm $\log_{10}$ of the relative energy error from endtimepoint $t = 0$ ▷

Observation:

$h\omega$-Dependence on critical frequencies

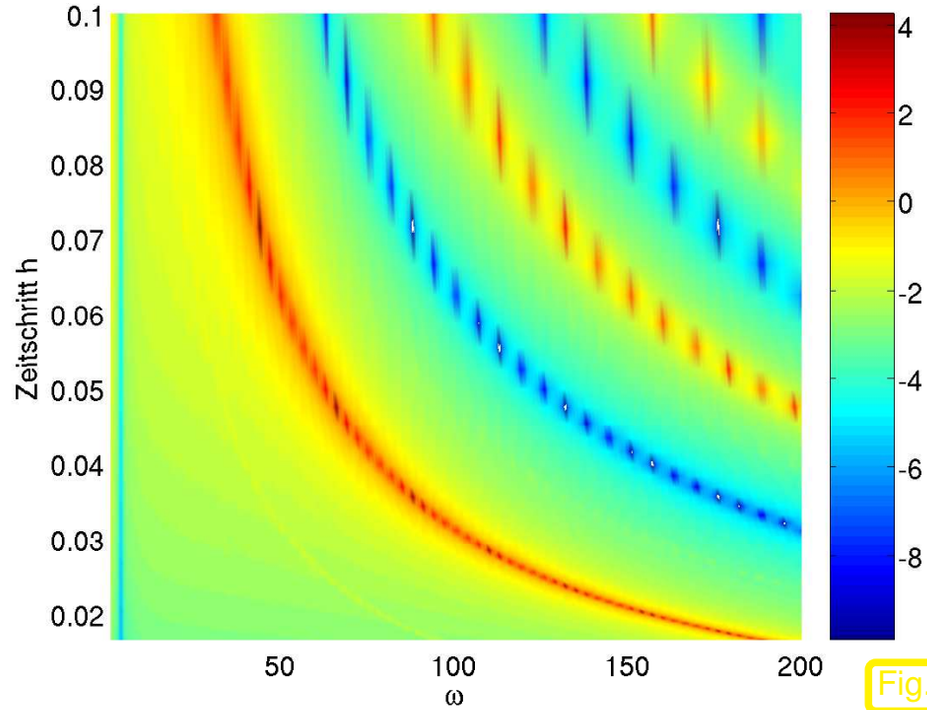


Fig. 203

◇

P. Grohs

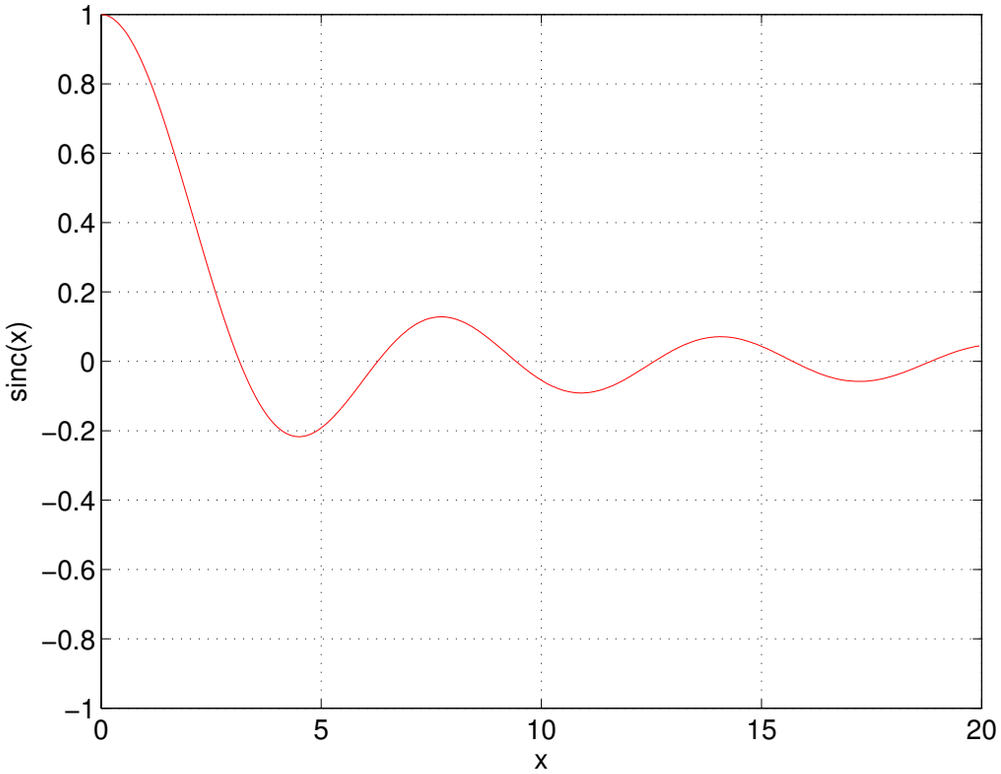
rev 63606, February 17, 2014

Model problem:

$$\ddot{y} = -\omega^2 y + \alpha y \ , \quad \alpha \ll \omega^2 \ . \tag{4.5.15}$$

➤ Gautschi-method (4.5.10), stepwidth h: ➢ Three-term-recursion

$$y_h(t+h) - \left\{ 2\cos(h\omega) + h^2 \alpha \operatorname{sinc}^2(\tfrac{1}{2}h\omega) \right\} y_h(t) + y_h(t-h) = 0 \ . \tag{4.5.16}$$



The sinc function:

$$\operatorname{sinc}(x) := \frac{\sin x}{x}$$

- Analytic on $\mathbb{R}$
- $|\operatorname{sinc}(x)| \leq 1$ mit global maximum in $x = 0$

P. Grohs

rev 63606, February 17, 2014

Analysis of (4.5.16): Ansatz $y_h(kh) = \xi^k$ $\leftrightarrow$ Characteristic (quadratic) equation

$\exists$ Solutions $y_h(kh)$ of (4.5.16): $\lim\limits_{k\to\infty} y_h(hk) = \pm\infty \quad \Leftrightarrow \quad \left|2\cos(h\omega) + h^2\alpha\operatorname{sinc}^2(\tfrac{1}{2}h\omega)\right| > 2$

$\blacktriangleright$ $(\cos h\omega \approx 1 \quad \Leftrightarrow \quad h\omega \approx 2\pi l,\ l \in \mathbb{Z}) \quad \Rightarrow \quad y_h(hk) \to \pm\infty$ auch für $h \ll 1$.

Corrective [28]: "“Filtering”: Absorption of α, if $h\omega \approx 2\pi l$:

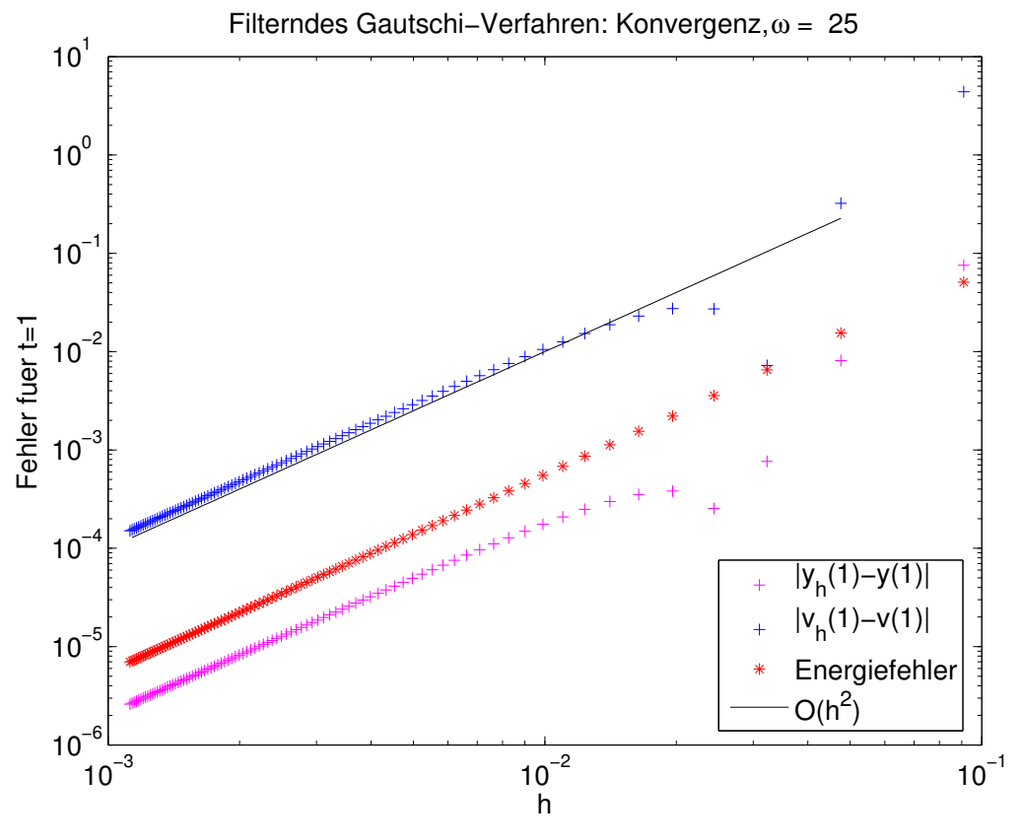
In (4.5.10), (4.5.11) replace: $g(y_h(t)) \quad \mapsto \quad g(\psi(h\omega)y_h(t)), \quad \psi(\xi) := \operatorname{sinc}^2 \xi(1 + \frac{1}{2}(1 - \cos \xi))$

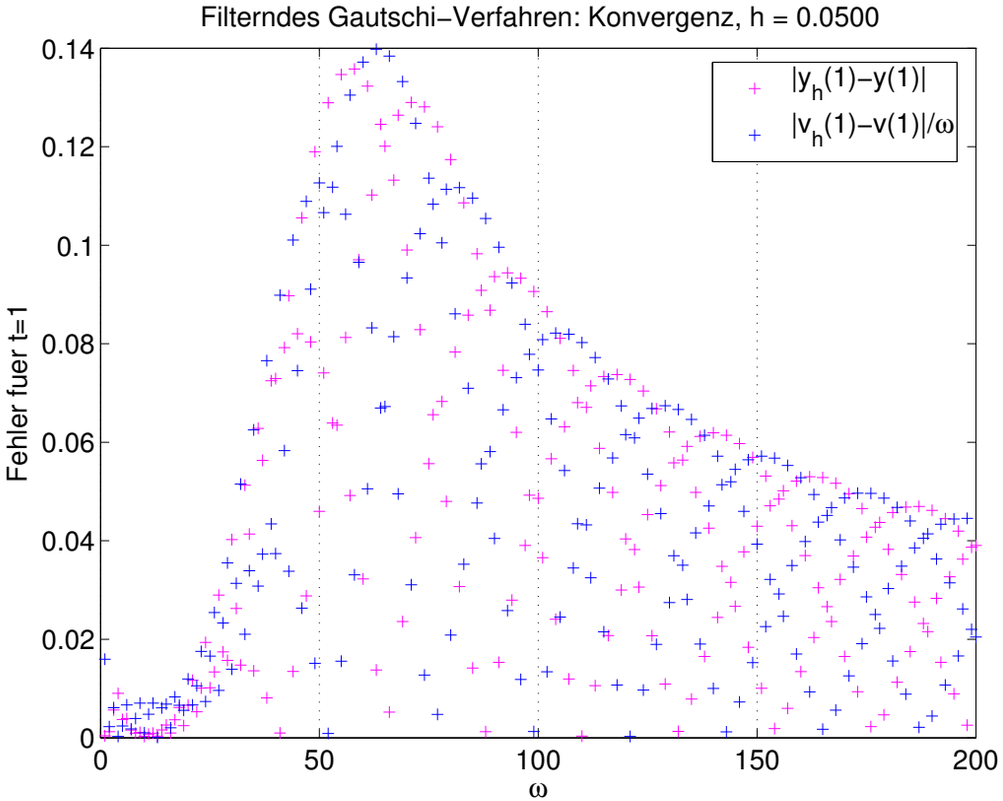
➤ modified gautschi-method:

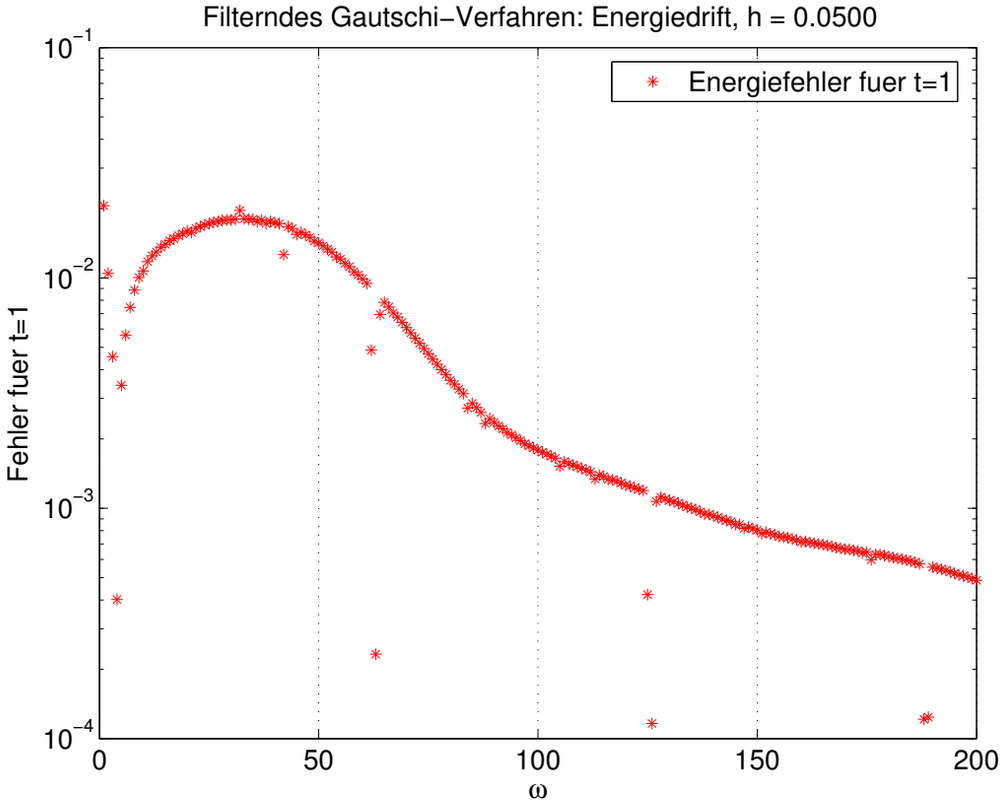
$$y_h(t+h) - 2\cos(h\omega)y_h(t) + y_h(t-h) = h^2 \left(\frac{\sin(\frac{1}{2}h\omega)}{\frac{1}{2}h\omega} \right)^2 g(\psi(h\omega)y_h(t)) \ . \tag{4.5.17}$$

Example 4.5.18 (Modified Gautschi-method).
IVP from Ex. 4.5.10, integration according to (4.5.17), filter function $\psi(\xi) := \operatorname{sinc}^2 \xi(1 + \frac{1}{2}(1 - \cos \xi))$

Convergence order 2 ▷



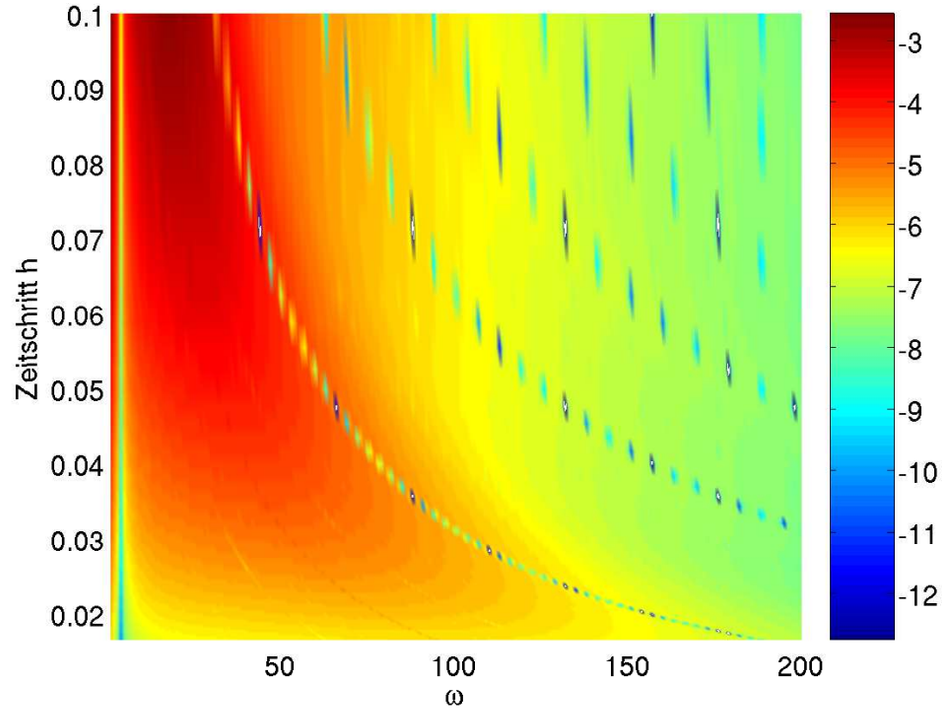

Filterndes Gautschi−Verfahren: Konvergenz, h = 0.0500
|y_h(1)−y(1)|
|v_h(1)−v(1)|/ω
Fehler fuer t=1
ω
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0
50
100
150
200


Filterndes Gautschi−Verfahren: Energiedrift, h = 0.0500
Energiefehler fuer t=1
Fehler fuer t=1
ω
10^−1
10^−2
10^−3
10^−4
0
50
100
150
200

$h\omega$-dependence of the energy drift ▷

Observation:

(4.5.17):
- No instability
- In comparison to (4.5.10) (→ Ex. 4.5.14) significantly reduced energy drift (Scale!)



P. Grohs

rev 63606, February 17, 2014

Index

P. Grohs

rev 63606, February 17, 2014

Index

P. Grohs

rev 63606, February 17, 2014

Appendix

MATLAB-Files for Examples

- Example 1.2.5 $\leftrightarrow$ File/Directory `ex:LV`
- Example 1.2.12 $\leftrightarrow$ File/Directory `ex:Oregonator`
- Example 1.2.15 $\leftrightarrow$ File/Directory `ex:heartbeat`
- Example 1.3.36 $\leftrightarrow$ File/Directory `ex:Lorenz`
- Example 1.4.4 $\leftrightarrow$ File/Directory `ex:expleulcvg`
- Example 1.4.9 $\leftrightarrow$ File/Directory `ex:eeullog`

- Example 1.4.15 $\leftrightarrow$ File/Directory `ex:ieullog`
- Example 1.4.17 $\leftrightarrow$ File/Directory `ex:pendeul`
- Example 1.4.18 $\leftrightarrow$ File/Directory `ex:eulspin`
- Example 1.4.21 $\leftrightarrow$ File/Directory `ex:logimid`
- Example 1.4.22 $\leftrightarrow$ File/Directory `ex:imidspin`

```
>> eulspin([1;0],10,40,'midspin40')
>> eulspin([1;0],10,160,'mispin160')
```

- Example 1.4.24 $\leftrightarrow$ File/Directory `ex:pendimid`

```
>> pendmidp([pi/4;0],5,50,'pendimid50');
>> pendmidp([pi/4;0],5,100,'pendimid100');
>> pendmidp([pi/4;0],5,200,'pendimid200');
```

- Example 1.4.32 $\leftrightarrow$ File/Directory `ex:svpend`
- Example 2.2.56 $\leftrightarrow$ File/Directory `ex:cvgkoll`
- Example 2.2.48 $\leftrightarrow$ File/Directory `ex:GaussCollcvg`
- Example 2.4.2 $\leftrightarrow$ File/Directory `ex:kombesv`
- Example 2.3.22 $\leftrightarrow$ File/Directory `ex:rkexplcvg`
- Example 2.4.17 $\leftrightarrow$ File/Directory `ex:eulexpol`
- Example 2.4.19 $\leftrightarrow$ File/Directory `ex:eulex`

- Example 2.5.4 ↔ File/Directory `ex:splitcvg`
- Example 2.5.14 ↔ File/Directory `ex:splitinex`
- Example 2.6.4 ↔ File/Directory `ex:qualest`
- Example 2.6.15 ↔ File/Directory `ex:adesv`
- Example 2.6.16 ↔ File/Directory `ex:adaptsat`
- Example 3.0.1 ↔ File/Directory `ex:logeximpl`
- Example 3.3.14 ↔ File/Directory `ex:GaussCollLog`
- Example 3.4.1 ↔ File/Directory `ex:iesvstiff`
- Example 3.5.2 ↔ File/Directory `ex:ode45stiff`
- Example 3.5.5 ↔ File/Directory `ex:odecircuit`
- Example 3.5.7 ↔ File/Directory `ex:odes`
- Example 3.6.1 ↔ File/Directory `ex:silog`

```
logsieul(0.1,round(5*1.5.^(0:18)),'silog');
```

- Example 3.6.3 ↔ File/Directory `ex:siradlog`

```
siradlog(0.1,round(5*1.5.^(0:18)),'siradlog')
```

- Example 3.6.12 ↔ File/Directory `ex:pendimipEnSmp`
- Example 3.7.5 ↔ File/Directory `ex:eeul`

- Example 3.7.6 $\leftrightarrow$ File/Directory `ex:eeulstiff`
- Example 3.8.15 $\leftrightarrow$ File/Directory `ex:daecircml`
- Example 3.8.28 $\leftrightarrow$ File/Directory `ex:daependmatlab`
- Example 3.8.29 $\leftrightarrow$ File/Directory `ex:daependieul`
- Example 4.1.3 $\leftrightarrow$ File/Directory `ex:magneedle`
- Example 4.4.1 $\leftrightarrow$ File/Directory `ex:enpres`
- Example 4.4.33 $\leftrightarrow$ File/Directory `ex:pendsympeul`
- Example 4.4.37 $\leftrightarrow$ File/Directory `ex:md`
- Example 4.4.35 $\leftrightarrow$ File/Directory `ex:springpend`
- Example 4.4.39 $\leftrightarrow$ File/Directory `ex:moldyn`

Bibliography

[1] A. AL-MOHY AND N. HIGHAM, *Computing the action of the matrix exponential with an application to exponential integrators*, MIMS ePrint 2010.30, Manchester Institute for Mathematical Sciences. University of Manchester, Manchester, UK, 2010.

[2] H. AMANN, *Gewöhnliche Differentialgleichungen*, Walter de Gruyter, Berlin, 1st ed., 1983.

[3] V. ARNOLD, *Mathematical Methods of Classical Mechanics*, Springer, New York, 2nd ed., 1989.

[4] L. BANJAI, C. LUBICH, AND J. MELENK, *Runge-kutta convolution quadrature for operators arising in wave propagation*, Numer. Math., 119 (2011), pp. 1–20.

[5] C. BLATTER, *Analysis I*, vorlesungsskriptum, ETH Zürich, Zürich, Switzerland, 2003. http://www.math.ethz.ch/~blatter/.

[6] L. BRUGNANO, F. IAVERNARO, AND D. TRIGIANTE, *Hamiltonian boundary value methods (energy preserving discrete line integral methods)*, JNAIAM. J. Numer. Anal. Ind. Appl. Math., 5 (2011), pp. 17–37.

[7] M. CHAWLA AND M. JAIN, *Error estimates for gauss quadrature formulas for analytic functions*, Math. Comp., 22 (1968), pp. 82–90.

[8] P. DAVIS, *Interpolation and Approximation*, Dover, New York, 1975.

[9] M. DEAKIN, *Applied catastrophe theory in the social and biological sciences*, Bulletin of Mathematical Biology, 42 (1980), pp. 647–679.

[10] P. DEUFLHARD, *Recent progress in extrapolation methods for ordinary differential equations*, SIAM Rev., 27 (1985), pp. 505–535.

[11] P. DEUFLHARD, *Numerik von anfangswertaufgaben für gewöhnliche differentialgleichungen*, Tech. Rep. TR 89-2, ZIB Berlin, Berlin, Germany, 1989.

[12] P. DEUFLHARD AND F. BORNEMANN, *Scientific Computing with Ordinary Differential Equations*, vol. 42 of Texts in Applied Mathematics, Springer, New York, 2 ed., 2002.

[13] P. DEUFLHARD AND A. HOHMANN, *Numerische Mathematik I*, DeGruyter, Berlin, 3 ed., 2002.

[14] G. FISCHER, *Lineare Algebra*, Vieweg–Verlag, Braunschweig, 9th ed., 1986.

[15] C. GRAY, *An analysis of the Belousov-Zhabotinski reaction*, Rose-Hulman Undergraduate Math Journal, 3 (2002). http://www.rose-hulman.edu/mathjournal/archives/2002/vol3-n1/paper1/v3n1-1pd.pdf.

[16] W. HACKBUSCH, *Iterative Lösung großer linearer Gleichungssysteme*, B.G. Teubner–Verlag, Stuttgart, 1991.

[17] E. HAIRER, *Energy-preserving variant of collocation methods*, JNAIAM. J. Numer. Anal. Ind. Appl. Math., 5 (2011), pp. 73–84.

[18] E. HAIRER AND C. LUBICH, *Asymptotic expansions and backward analysis for numerical integrators*, in Dynamics of algorithms, vol. 118 of IMA Vol. Math. Appl., Springer, New York, 2000, pp. 91–106.

[19] E. HAIRER, C. LUBICH, AND G. WANNER, *Geometric Numerical Integration*, vol. 31 of Springer Series in Computational Mathematics, Springer, Berlin, Germany, 2002. Table of contents.

[20] ——, *Geometric numerical integration illustrated by the Störmer-Verlet method*, Acta Numerica, 12 (2003), pp. 399–450.

[21] ——, *Geometric numerical integration*, vol. 31 of Springer Series in Computational Mathematics, Springer, Heidelberg, 2 ed., 2006.

[22] E. HAIRER, S. NORSETT, AND G. WANNER, *Solving Ordinary Differential Equations I. Nonstiff Problems*, Springer-Verlag, Berlin, Heidelberg, New York, 2 ed., 1993.

[23] E. HAIRER AND G. WANNER, *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*, vol. 14 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 2011.

[24] M. HERRMANN, *Numerik gewöhnlicher Differentialgleichungen*, Oldenbourg, München, 2004.

[25] N. HIGHAM, *The scaling and squaring method for the matrix exponential revisited*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 1179–1193.

[26] M. HIRSCH, S. SMALE, AND R. DEVANEY, *Differential Equations, Dynamical Systems, and an Introduction to Chaos*, vol. 60 of Pure and Applied Mathematics, Elsevier Academic Press, Amsterdam, 2 ed., 2004.

[27] M. HOCHBRUCK AND C. LUBICH, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925.

[28] —, *A Gautschi-type method for oscillatory second-order differential equations*, Numer. Math., 83 (1999), pp. 403–426.

[29] M. HOCHBRUCK, C. LUBICH, AND H. SELHOFER, *Exponential integrators for large systems of differential equations*, SIAM J. Sci. Comp., 19 (1998), pp. 1552–1574.

[30] M. HOCHBRUCK AND A. OSTERMANN, *Exponential integrators*, Acta Numerica, 19 (2010), pp. 209–286.

[31] B. LEIMKUHLER AND S. REICH, *Simulating Hamiltonian Dynamics*, vol. 14 of Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, UK, 2004.

[32] E. LORENZ, *Deterministic non-periodic flow*, J. Atmospheric Sciences, 20 (1963), pp. 130–141.

[33] B. MINCHEV AND W. WRIGHT, *A review of exponential integrators for first order semi-linear problems*, Preprint 2/2005, NORGES TEKNISK-NATURVITENSKAPELIGE UNIVERSITET, Trondheim, Norway, 2005.

[34] H. OGATA, *A numerical integration formula based on the Bessel functions*, Publ. Res. Inst. Math. Sci., 41 (2005), pp. 949–970.

[35] S. REICH, *Backward error analysis for numerical integrators*, SIAM J. Numer. Anal., 36 (1999), pp. 1549–1570.

[36] R. REMMERT, *Funktionentheorie I*, no. 5 in Grundwissen Mathematik, Springer, Berlin, 1984.

[37] L. SHAMPINE, M. REICHELT, AND J. KIERZENKA, *The MATLAB ODE suite*, SIAM J. Sci. Comp., 18 (1997), pp. 1–22.

[38] W. WALTER, *Gewöhnliche Differentialgleichungen. Eine Einführung*, vol. 110 of Heidelberger Taschenbücher, Springer, Heidelberg, 3 ed., 1986.